Pontifícia Universidade Católica de São Paulo
PUC-SP

Aline Tomasuolo Souza

É possível distinguir a tradução automática da tradução humana? Uma perspectiva baseada em corpus e aprendizagem de máquina

Mestrado em Linguística Aplicada e Estudos da Linguagem

São Paulo
2023

Aline Tomasuolo Souza

É possível distinguir a tradução automática da tradução humana? Uma perspectiva baseada em corpus e aprendizagem de máquina

Mestrado em Linguística Aplicada e Estudos da Linguagem

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRA em Linguística Aplicada e Estudos da Linguagem, sob a orientação do Prof. Dr. Antonio Paulo Berber Sardinha.

São Paulo

2023

Aline Tomasuolo Souza

É possível distinguir a tradução automática da tradução humana? Uma perspectiva baseada em corpus e aprendizagem de máquina

Aprovada em: ____/____/_____

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Linguística Aplicada e Estudos da Linguagem, sob orientação do Professor Doutor Antonio Paulo Berber Sardinha,

Banca Examinadora:

_____

Prof. Dr. Antonio Paulo Berber Sardinha – Orientador

_____

Prof. Dra. Elaine Alves Trindade

_____

Prof. Dra. Marilisa Shimazumi

Ao meu filho Henrique e ao meu parceiro de vida Marcio, amores da minha vida, daqui até a eternidade, à minha mãe Ana e minhas irmãs Bruna e Luiza, minhas obstinadas companheiras, pelo amor e apoio incondicionais ao longo dessa jornada.

Agradecimento à CAPES

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Número do processo: 88887.473563/2020-00
Período: 01/02/2020 a 31/10/2022

Agradecimentos

Em primeiro lugar, gostaria de agradecer, pelos grandes aprendizados proporcionados, ao Programa de Pós-Graduação em Linguística Aplicada e Estudos da Linguagem da Pontifícia Universidade Católica de São Paulo, e à CAPES, por viabilizar essa oportunidade ímpar de ampliar os horizontes do meu conhecimento acadêmico e profissional.

Agradeço imensamente a cada um dos professores com quem aprendi tanto ao longo desses três anos: Prof. Dra. Sumiko Nishitani Ikeda, Prof. Dra. Mara Sofia de Toledo Zanotto, Prof. Dra. Maximina Maria Freire e Prof. Dra. Maria Cecília Pérez de Souza e Silva. Além dos muitos conhecimentos transmitidos por essas renomadas professoras, em todas as disciplinas que cursei a humanidade sempre imperou em sala de aula, para além do mundo acadêmico – o que foi essencial, devido ao momento delicado que todos vivenciamos ao longo da pandemia, e a um momento particular e especial da minha vida pessoal, que inclui a gestação, puerpério e maternagem do meu incrível filho.

No entanto, academicamente, o maior agradecimento de todos pertence ao meu orientador, Prof. Dr. Tony Berber Sardinha. Em minha modesta opinião, o professor Tony é um gênio de nossa geração como estudioso da linguística e ficará marcado para sempre na história da Linguística de Corpus no Brasil. Ele combina características inestimáveis para um professor, transmitindo conhecimento de forma didática e inovadora, e características impressionantes como estudioso e acadêmico, em busca das mais novas tecnologias e sua aplicabilidade ao mundo da linguística, não só descobrindo novas ferramentas, como criando novos métodos, recursos e usos para tais.

Por fim, minha gratidão eterna à minha família. Ao meu filho Henrique, que começou essa jornada comigo em meu ventre e está prestes a completar três anos de idade. Toda minha força, coragem e garra vem dele, para ele. Ao meu parceiro de vida, Marcio – sem você, não teria chegado até aqui. Sem seu apoio inabalável, seu amor e sua ‘abraçoterapia’, nada disso seria possível. À minha mãe, guerreira e leoa – seu amor incondicional me trouxe até aqui. À minha irmã Bruna, que sempre me deu forças para não desistir, e à minha irmã Luiza, que sempre foi minha cúmplice e confidente.

Obrigada!

“O que veio primeiro, a fênix ou a chama?”

J.K. Rowling (Tradutora: Lia Wyler)

Resumo

Nos últimos anos, houve avanços significativos nas tecnologias de tradução automática, levando ao questionamento sobre sua eficácia em relação à tradução humana. Nesta dissertação de mestrado, exploramos essa questão por meio de uma abordagem baseada em corpus e aprendizagem de máquina. O corpus compilado inclui textos em inglês da área financeira, especificamente de companhias de capital aberto, sendo textos provavelmente produzidos em inglês por nativos do idioma e textos traduzidos do português para o inglês. O corpus foi dividido em três subcorpora: corpus de textos em inglês provavelmente produzidos por nativos do idioma (corpus comparável), corpus de tradução humana e corpus de tradução automática (corpora paralelos). Utilizamos o Biber Tagger para análise gramatical e o Weka para análise lexical dos corpora. Com o Biber Tagger, examinamos as estruturas gramaticais do corpus. Por meio do Weka, realizamos uma análise lexical nos corpora, identificando diferenças e semelhanças entre a tradução automática, a tradução humana e textos provavelmente escritos por nativos da língua inglesa. Esta abordagem nos permitiu criar um modelo probabilístico que pode prever, com 85% de precisão, se uma tradução foi produzida por uma máquina ou um tradutor humano. Assim, concluímos que, lexicalmente, é possível diferenciar a tradução automática da tradução humana; no entanto, gramaticalmente, ambas as traduções são equiparáveis e em níveis comparáveis aos textos provavelmente escritos por falantes nativos de inglês.

Palavras-chave: Tradução Automática; Linguística de Corpus; Análise Multidimensional; Análise Lexical.

Abstract

In recent years, there have been significant advances in machine translation technologies, leading to questions about their effectiveness compared to human translation. In this master's dissertation, we explore this issue through a corpus-based and machine-learning approach. The compiled corpus includes English texts from the financial area, specifically from listed companies, including translated texts from Portuguese to English and texts written in English by native speakers. The corpus was divided into three subcorpora: an English-native text corpus (comparable corpus), a human translation corpus, and an automatic translation corpus (parallel corpora). We used the Biber Tagger for grammatical analysis and Weka for lexical analysis of the corpora. With the Biber Tagger, we examined the grammatical structures of the corpus. Through Weka, we conducted a lexical analysis of the corpora, identifying differences and similarities between automatic translation, human translation, and texts written by native English speakers. This approach allowed us to create a probabilistic model that can predict, with 85% accuracy, if a translation was produced by a machine or a human translator. We concluded that lexically, it is possible to differentiate automatic translation from human translation; however, grammatically, both translations are nearly identical and at comparable levels to texts written by native English speakers.

Keywords: Machine Translation; Corpus Linguistics; Multidimensional Analysis; Lexical Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AA	Comunicado/Aviso aos Acionistas/Detentores de Títulos de Crédito/Detentores de Debêntures
ABRATES	Associação Brasileira de Tradutores e Intérpretes
ALPS	Automated Language Processing Systems
AT	Atas das Assembleias (Extraordinária/Ordinária)
ATA	American Translators Association
BNC	British National Corpus
CAT	Computer Assisted Translation
CM	Comunicado/Aviso ao Mercado
COCA	Corpus of Contemporary American English
CS	Comparable Subcorpus
CVM	Comissão de Valores Mobiliários
EAMT	European Association for Machine Translation
ENG-TM	Textos e/ou corpus em língua inglesa a partir de uma tradução automática
ENG-TRAD	Textos e/ou corpus em língua inglesa a partir de uma tradução humana-oficial
ENG-ORIG	Textos e/ou corpus em língua inglesa provavelmente produzidos por nativos do idioma
ES	Estatuto Social
FR	Fato Relevante
GAAP	Generally Accepted Accounting Principles
GNMT	Google Neural Machine Translation
GTPS	Google Translate Parallel Subcorpus
HCPS	Human Certified Parallel Subcorpus
IA	Inteligência Artificial
IFRS	International Financial Reporting Standards
LLC	London Lund Corpus of Spoken English
LOB	Lancaster-Oslo/Bergen Corpus
ML	Aprendizagem por Máquina
MT	Tradução Automática
NLP	Natural Language Processing
NMT	Tradução Automática Neural
NOW	News on the Web
PLN	Processamento de Linguagem Natural
PO	Políticas
PTBR-ORIG	Textos e/ou corpus em língua portuguesa produzido por nativos do idioma
SEC	Securities and Exchange Commission
TM	Memórias de Tradução
TSB	Translation Studies Bibliography
Weka	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 Introdução

Ao longo dos últimos anos, testemunhamos avanços extraordinários nas tecnologias de tradução automática. Com o desenvolvimento de algoritmos sofisticados e o poder crescente da aprendizagem de máquina, a tradução automática se tornou uma ferramenta amplamente utilizada em várias áreas, desde comunicações internacionais até traduções sofisticadas de áreas complexas.

Isso provocou um questionamento: será que chegamos ao ponto em que a tradução automática é tão eficaz e precisa quanto a tradução humana? Nesta dissertação de mestrado, exploramos essa pergunta intrigante, adotando uma perspectiva inovadora baseada em corpus e aprendizagem de máquina.

A tradução automática, impulsionada por avanços na inteligência artificial, tem se mostrado cada vez melhor em produzir traduções fluentes e coerentes. No entanto, apesar dos esforços para aprimorar a qualidade da tradução automática, muitos linguistas e tradutores profissionais ainda questionam se ela pode alcançar o mesmo nível de qualidade e sutileza que a tradução humana.

Essa discussão levanta importantes considerações sobre a criatividade, a compreensão cultural e a habilidade de tomar decisões contextuais, características que há muito tempo se acredita serem exclusivas do tradutor humano. Enquanto as tecnologias de tradução continuam a avançar, torna-se cada vez mais importante compreender as diferenças sutis e as similaridades entre a tradução automática e a humana.

A atual pesquisa teve como base dois grandes questionamentos iniciais, sendo: 1) Há traços linguísticos que podem diferenciar de forma probabilística a tradução humana da tradução automática?; e 2) Há traços linguísticos que podem diferenciar de forma probabilística e estatística uma tradução para o inglês (seja ela automática ou humana) de um texto produzido originalmente em inglês?

Para isso, o atual estudo usou como base teórico-metodológica a Linguística de Corpus, área dos Estudos Linguísticos ocupada na análise de grandes porções de textos que representam um dado domínio ou registro (BERBER SARDINHA, 2004), auxiliado por técnicas da linguística computacional, linguística aplicada, ciência da computação e inteligência artificial, combinadas para trazer novas perspectivas sobre a tradução (MANNING & SCHUTZE, 1999).

Ao analisar grandes volumes de textos traduzidos por humanos e por sistemas automáticos, o objetivo é identificar padrões linguísticos e características distintivas que podem revelar indícios sobre a origem da tradução. Essa abordagem inovadora nos permite explorar

até que ponto a tradução automática pode replicar a complexidade e a qualidade da tradução humana, levando em consideração aspectos gramaticais e lexicais. Como será visto a seguir, através do treinamento de algoritmos em grandes conjuntos de dados de traduções humanas e automáticas, buscou-se criar modelos preditivos capazes de distinguir com razoável precisão, do ponto de vista lexical, uma tradução gerada por máquina de uma tradução realizada por um tradutor humano.

Conforme detalhado no Capítulo 2, a Linguística de Corpus envolve o exame sistemático e quantitativo de amostras de linguagem real (MCENERY, XIAO & TONO, 2006), de modo a ser usada para estatisticamente analisar as diferenças entre a tradução automática e a tradução humana e gerar um modelo probabilístico. Com base nesses dados, foi possível identificar padrões ou características distintivas que diferenciam os dois tipos de tradução. Compreender até que ponto a tradução automática pode substituir a tradução humana pode influenciar as decisões sobre o uso dessas tecnologias em diversos contextos, como tradução de documentos oficiais, comunicação intercultural e até mesmo na literatura e nas artes, além de potencialmente gerar implicações significativas para a indústria da tradução, bem como para os campos da inteligência artificial e da linguística computacional.

Para alcançar tais resultados, conforme detalhado no Capítulo 3, compilamos um corpus com 3.262.082 palavras. Esse corpus consiste em textos corporativos em inglês, especificamente da área financeira, de companhias de capital aberto (com ações listadas na bolsa de valores brasileira e na bolsa de valores de Nova York), incluindo comunicados e documentos de governança corporativa.

O corpus dessa pesquisa se divide em três subcorpora:

1. Corpus comparável: totalizando 867.446 palavras, esse corpus foi coletado dos sites de relações com investidores de companhias americanas. Sendo assim, são textos em inglês provavelmente produzidos por falantes nativos da língua inglesa. Esse corpus foi utilizado para investigar se há diferenças gramaticais e lexicais entre textos provavelmente escritos por falantes nativos e textos traduzidos.

2. Corpus paralelo de tradução humana: totalizando 764.895 palavras, esse corpus foi coletado da versão em inglês dos sites de relações com investidores de companhias brasileiras. Tais textos em inglês são versões, ou textos produzidos originalmente em português traduzidos para o inglês.

3. Corpus paralelo de tradução automática: totalizando 813.997 palavras, esse corpus teve uma coleta um pouco mais complexa. Coletamos os textos originais em português, equivalentes as versões em inglês coletadas para o corpus paralelo de tradução humana, dos

mesmos sites de relações com investidores. Tais textos em português foram traduzidos do português para o inglês por uma ferramenta de tradução automática (Google) para assim compilarmos o corpus de tradução automática.

Para realizar a análise gramatical dos corpora utilizados nesta pesquisa, utilizamos o Biber Tagger (BIBER, 1988), um etiquetador para língua inglesa que é uma ferramenta reconhecida no campo da Linguística de Corpus. O Biber Tagger é um sistema de marcação morfossintática que utiliza técnicas de processamento de linguagem natural para atribuir etiquetas gramaticais a cada palavra do corpus. Com essa abordagem, examinamos e comparamos as estruturas linguísticas dos três subcorpora, identificando diferenças (ou ausência destas) e avaliando a qualidade de cada tipo de tradução, conforme descrito no Capítulo 4.

Para a análise lexical dos corpora, utilizamos a ferramenta Weka, uma plataforma de aprendizado de máquina que oferece uma variedade de algoritmos e ferramentas para a mineração de dados textuais (WITTEN et al., 2017). Com o Weka, foi possível realizar uma análise das palavras-chave, frequências e padrões vocabulares encontrados nos corpora de traduções automáticas e humanas. Isso nos permitiu explorar diferenças e semelhanças entre os dois tipos de tradução em termos de escolhas lexicais, avaliando a riqueza vocabular e a adequação do uso de termos em cada contexto. Além disso, com o auxílio dessa ferramenta, será mostrado como foi possível criar um modelo probabilístico capaz de classificar com precisão se uma tradução testada era uma tradução automática ou uma tradução humana. Ao combinar a análise gramatical proporcionada pelo Biber Tagger e a análise lexical facilitada pelo Weka, ampliamos nossa compreensão das características, similaridades e distinções entre os corpora paralelos e o corpus comparável.

Por fim, conforme detalhado no Capítulo 5, através dessa abordagem com base na Linguística de Corpus e aprendizagem de máquina, será apresentado como foi possível atingir os objetivos desta pesquisa e responder à pergunta do título, à guisa de conclusão.

2 Fundamentação teórica

Esta pesquisa foi fundamentada com base em duas áreas distintas de estudo, a Linguística de Corpus e os Estudos de Tradução. Essas duas áreas foram relacionadas e conectadas para investigar e explorar os efeitos da revolução tecnológica na área de tradução com base na Linguística de Corpus.

Sendo assim, a fundamentação teórica apresentará individualmente cada uma dessas áreas visando esclarecer e respaldar a pesquisa com o devido suporte teórico.

2.1 Linguística de Corpus

2.1.1 Histórico e perspectivas

O surgimento da Linguística de Corpus como área de pesquisa da linguística aplicada está intrinsecamente ligado aos avanços tecnológicos na área de computação e processamento de dados. Seus primeiros passos podem ser situados no começo do século XX, quando pesquisadores começaram um processo para compilar um conjunto de textos para analisá-los sob o ponto de vista linguístico. No entanto, a falta de ferramentas computacionais e o tempo considerável para analisar manualmente os textos compilados foram desafiadores, impondo limitações ao alcance desse estudo (BERBER SARDINHA, 2004; SVARTVIK, 1990).

Os anos 1960 foram essenciais para a Linguística de Corpus como conhecemos hoje. O primeiro corpus linguístico eletrônico foi desenvolvido na Brown University em 1964 (FRANCIS & KUČERA, 1979), inovando a área de estudos linguísticos. Para efeito de comparação, o primeiro computador pessoal e a internet só foram desenvolvidos 10 anos depois, em 1974. Os avanços tecnológicos de então, na área de informática, permitiram que a Brown University pudesse, de forma eficiente, armazenar, processar e analisar um volume grande de dados, para aquela época, a partir dos textos compilados.

As tecnologias computacionais continuaram a evoluir ao longo dos anos 70 e 80 do século XX, o que beneficiou a Linguística de Corpus e permitiu o desenvolvimento de ferramentas computacionais cada vez mais sofisticadas e avançadas. Sendo assim, pesquisadores da área começaram a realizar análises mais complexas destes conjuntos de dados e, consequentemente, descobrir e investigar padrões, colocações e frequências (STUBBS, 1995). Nessa época tivemos também a criação de corpora importantes para a área, como o Corpus Lancaster-Oslo/Bergen (LOB - *Lancaster-Oslo/Bergen Corpus*) (JOHANSSON, 1978)

e o Corpus London-Lund do Inglês Falado (LLC - *London Lund Corpus of Spoken English*) (SVARTVIK, 1990).

Nos anos 1990, a criação do Corpus Nacional Britânico (BNC - *British National Corpus*) foi um marco significativo (BURNARD, 1995). O BNC, na sua primeira versão, é um corpus com 100 milhões de palavras, com diversas variantes de textos em inglês britânico, escritos e falados.

A partir disso, deu-se início à criação de outros corpora de tamanhos cada vez mais consideráveis, como o Corpus do Inglês Americano Contemporâneo (COCA – *Corpus of Contemporary American English*), atualmente com cerca de 1 bilhão de palavras, e o *News on the Web (*NOW), atualmente com mais de 17 bilhões de palavras. Na imagem abaixo, extraída do site www.english-corpora.org, pode-se observar o tamanho, idioma, período e gênero de alguns dos maiores corpora em língua inglesa atuais.

Figura 1 – Composição de alguns corpora em inglês de grandes dimensões

Corpus Overview	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)		17.5 billion+	20 countries	2010-yesterday	Web: News
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus		1.5 billion	20 countries	Jan 2020-Dec 2022	Web: News
Corpus of Contemporary American English (COCA)		1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)		475 million	American	1820-2019	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 billion	British	1803-2005	Parliament
Early English Books Online		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1920s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web
From Google Books n-grams (compare)					
American English		155 billion	American	1500s-2000s	(Various)
British English		34 billion	British	1500s-2000	(Various)

Fonte: https://www.english-corpora.org/

Com a construção de corpora em outras línguas além da língua inglesa, pesquisadores começaram a desenvolver corpora paralelos, que são corpora com versões de um mesmo texto

em duas ou mais línguas (BAKER, 2000; RESENDE, 2019), como explicado abaixo no Capítulo 3. Tais corpora foram essenciais para a linguística comparativa, os estudos de tradução e o desenvolvimento de algoritmos de processamento da linguagem natural. Um dos corpus criados nesse âmbito foi o Corpus Europarl, com procedimentos e casos do Parlamento Europeu, em 21 idiomas (KOEHN, 2005).

A internet teve um grande impacto na Linguística de Corpus. A disponibilização de diversos textos digitais de diversas fontes permitiu criar corpora com base em textos online, tais como o WebCorp (RENOUF, KEHOE & BANERJEE, 2007), e estudos que tiveram como fonte dados disponíveis em plataformas online, como o Google Books Ngram Viewer (BERBER SARDINHA, 2020; MICHEL et al., 2011). Tal disponibilidade permitiu que estudiosos da área fizessem pesquisas com uma escala muito maior, identificando variações linguísticas ao longo dos anos com uma amplitude muito mais abrangente.

Ao longo de toda sua história, a Linguística de Corpus se consolidou como uma área importante para os estudos linguísticos e diversas são as suas aplicações na vida prática, na forma de dicionários (por exemplo, o *Collins COBUILD English Dictionary*), gramáticas (p. ex., o *Longman Grammar of Written and Spoken English*), publicações de referência e ferramentas de análise da linguagem (p. ex., a suíte de aplicativos Ant).

A área de Linguística de Corpus está cada vez mais interdisciplinar. A área de processamento da linguagem natural (NLP - *Natural Language Processing*), que permite a interação entre a linguagem computacional e a linguagem humana, é uma delas, por também trabalhar com dados linguísticos. A criação de corpora comentados de grande volume, tais como as bases *Penn Treebank* e *Universal Dependencies,* foi essencial para desenvolver algoritmos de aprendizagem por máquina utilizados em tarefas de NLP.

Nos Estados Unidos, Douglas Biber despontou como grande estudioso da área de Linguística de Corpus (BIBER, 1988, 1991, 1995; BIBER, CONRAD & REPPEN 1998; BIBER & CONRAD 2019). Em território brasileiro, Tony Berber Sardinha é referência, com inúmeras contribuições a essa área de estudo (BERBER SARDINHA, 2000, 2004; BERBER SARDINHA & VEIRANO PINTO 2014, 2019). Há quase 20 anos, no prefácio de seu livro "Linguística de Corpus", de 2004, Berber Sardinha fez uma previsão assertiva:

> Está em curso uma verdadeira revolução no pensamento linguístico, com implicações sérias sobre como respondemos a questões fundamentais, tais como o que é língua, como ela é organizada, como deve ser estudada, como deve ser ensinada. A mola propulsora dessa revolução é a tecnologia, mais especificamente o computador. (BERBER SARDINHA, 2004, p. 17).

2.1.2 Princípios e definições

Para entender a Linguística de Corpus, é preciso ter clareza em relação ao seu principal objeto de estudo: o corpus. Com a evolução tecnológica e semântica, 'corpus' já teve diversos significados diferentes com propósitos distintos. Sendo assim, considerando os objetivos desta pesquisa, a seguinte definição de corpus foi utilizada:

> "[...] um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos) sistematizados segundo determinados critérios suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de proporcionar resultados vários e úteis para a descrição e análise".
> (BERBER SARDINHA, 2004, p. 3).

A definição acima foi a mais representativa e corroborativa ao se considerar o objetivo desta pesquisa, que se resume a análise sistemática de dois corpora paralelos e um corpus comparável, englobando um conjunto de traduções automáticas e humanas e textos provavelmente escritos por falantes nativos do inglês, representativos da área financeira, processadas por computador, assim gerando diversos resultados, para responder à pergunta: há diferenças significativas e sistemáticas entre traduções realizadas por humanos e por máquinas?

A Linguística de Corpus é uma abordagem empírica que se baseia em dados observáveis e mensuráveis. Sendo assim, a Linguística de Corpus precisa se fundamentar no uso real da linguagem, em vez de teorias ou hipóteses. Os dados devem ser coletados e analisados de forma sistemática, por meio de ferramentas e técnicas da linguística, da estatística e da tecnologia computacional (BERBER SARDINHA, 2004).

Outro princípio da Linguística de Corpus é o uso de amostras representativas. O design do corpus é realizado de forma que o corpus seja uma amostra representativa de um idioma ou variedade específica de um idioma. Dessa forma, o resultado da análise do corpus permite generalizar as conclusões da análise para aquela população ou segmento específico. O tamanho e a composição do corpus podem variar de acordo com os objetivos da pesquisa e as características linguísticas analisadas. Um corpus, portanto, precisa ser composto de uma variedade de textos e fontes de forma a realmente representar a variabilidade linguística daquele idioma ou linguagem específica alvo de interesse, bem como das características linguísticas pesquisadas, tendo como base as perguntas de pesquisa (EGBERT, BIBER & GRAY, 2022).

A análise quantitativa também faz parte do rol de princípios da Linguística de Corpus, considerando que tal abordagem utiliza métodos quantitativos na análise de dados, tais como identificação de frequência, colocações e linhas de concordância. Isso faz com que seja possível identificar padrões e tendências nos dados e investigar hipóteses em relação ao uso da linguagem. Conforme Baker afirma:

> "*Corpus linguistics is firmly rooted in empirical, inductive forms of analysis, relying on real-world instances of language use in order to derive rules or explore trends about the ways in which people actually produce language [...]. A further advantage of the corpus linguistics approach is that it can enable researchers to quantify linguistic patterns, providing more solid conclusions*[1]."
> (BAKER, 2010, p. 94).

A Linguística de Corpus necessita que o analista leve em conta o contexto no qual os textos são inseridos: é preciso levar em conta o registro, o discurso, assim como fatores sociais e culturais da produção de tais textos. Ou seja, a Linguística de Corpus abrange a análise tanto qualitativa quanto quantitativa. A análise qualitativa do corpus, como explicado em mais detalhes acima, leva em conta a análise contextual dos dados, requisitando a interpretação dos textos.

Por fim, a Linguística de Corpus, como método de pesquisa no estudo da linguagem, tem se mostrado fonte de inúmeros estudos de diversos pesquisadores e universidades que compartilham conhecimentos da área e de outras disciplinas conexas, disponibilizando publicamente alguns dos corpora coletados e respectivos metadados obtidos em teses, dissertações, artigos, para encorajar e promover novas pesquisas e descobertas nesse campo de estudo.

2.1.3 Análise lexical na Linguística de Corpus

A análise lexical é fundamental para a Linguística de Corpus, com foco na investigação de palavras e padrões de uso no corpus por meio do estudo da frequência das palavras, distribuições, colocações e outras relações lexicais com o objetivo de compreender melhor o uso da linguagem, as variantes e as estruturas linguísticas subjacentes. A análise lexical permite

[1] Tradução da autora: A linguística de corpus está firmemente enraizada em formas empíricas e indutivas de análise, dependendo de instâncias do uso da linguagem no mundo real para derivar regras ou explorar tendências sobre as maneiras como as pessoas realmente produzem a linguagem [...]. Uma outra vantagem da abordagem da linguística de corpus é que ela pode permitir que os pesquisadores quantifiquem padrões linguísticos, fornecendo conclusões mais sólidas.

identificar padrões contextuais e terminológicos, sendo essencial para a linguística e estudos da tradução.

O entendimento destes padrões e estruturas lexicais permite discernir a natureza da linguagem e o funcionamento da linguagem em diferentes contextos. Com a análise sob a perspectiva lexical, pesquisadores identificam as palavras mais significativas e mais comuns em um texto, exploram os significados e conotações de tais palavras e descobrem padrões linguísticos menos óbvios.

Um dos principais elementos desse tipo de análise é o cálculo da frequência das palavras em um texto, que permite identificar quais palavras são usadas com mais frequência e quais palavras são menos frequentes. Esse dado pode revelar características importantes da linguagem estudada, tais como a terminologia específica utilizada nessa linguagem. Também é possível comparar essas características com outras variantes linguísticas e outros gêneros linguísticos e assim aprofundar ainda mais o conhecimento sobre as linguagens objeto de estudo e comparação.

Ao analisar a distribuição de palavras, é preciso considerar a dispersão das palavras no corpus. Essa análise pode revelar a variação no uso de palavras e coocorrências, tais como a tendência de usar certas palavras sempre em determinados textos e não em outros.

Uma das principais áreas de estudo da análise lexical é a análise de colocações. Colocações são pares de palavras que aparecem de forma frequente próximas uma da outra em um certo contexto, criando combinações significativas. A identificação de colocações em um corpus é essencial para desvendar como as palavras comumente se combinam e interagem de forma natural e, assim, estudar as estruturas adjacentes e convenções linguísticas que dão origem a tais combinações e interações. O conceito de Berber Sardinha foi utilizado para fundamentar a análise lexical referente às colocações no âmbito dessa pesquisa:

> "As colocações são um aspecto fundamental do uso da língua e são importantes para a compreensão do significado das palavras e dos textos. Elas são unidades lexicais que ocorrem frequentemente juntas na língua, e sua frequência de ocorrência pode ser medida por meio da análise de corpus. A análise de colocações envolve a identificação das palavras que ocorrem com maior frequência [...] é uma técnica importante na análise lexical de corpus e pode ser usada para a identificação de padrões linguísticos em diferentes gêneros e domínios de uso da língua."
> (BERBER SARDINHA, 2004, p. 73).

A concordância, ferramenta amplamente utilizada na Linguística de Corpus, é uma lista de ocorrências de uma palavra ou expressão específica em um corpus, juntamente com o contexto (texto próximo) relacionado a essas ocorrências. A análise de concordância permite

determinar como as palavras são usadas em diferentes contextos e explorar diferentes nuances de significados e associações.

O conceito de palavras-chave também é largamente empregado na Linguística de Corpus, definidas como palavras com frequência maior do que o esperado em um corpus específico quando comparado ao corpus de referência, indicando a saliência dessa palavra no corpus. As palavras-chave identificadas em um corpus podem ajudar a definir os principais temas, conceitos e pontos de atenção no texto e pode facilitar comparar as diferentes variantes ou gêneros textuais e/ou linguísticos.

Considerando todos os aspectos acima apresentados, a análise lexical é parte fundamental da Linguística de Corpus, levando a descobertas significativas sobre os padrões, estruturas e nuances de um idioma em uso. Ao examinar a frequência, a distribuição, as colocações, as linhas de concordância e as palavras-chave é possível desvendar os mecanismos e convenções linguísticos e assim aprofundar o conhecimento linguístico em diferentes contextos (BAKER, 2010).

2.1.4 Análise multidimensional funcional na Linguística de Corpus

A análise multidimensional é uma abordagem na Linguística de Corpus que estuda múltiplas características da linguagem de forma simultânea, permitindo a captura das interações entre essas características em diferentes dimensões. Cada dimensão representa um conjunto de características linguísticas que tendem a ocorrer juntas no texto, como frequência de palavras, estruturas gramaticais e padrões estilísticos. Ao examinar a prevalência dessas dimensões em um corpus, podemos obter insights sobre os padrões de uso da linguagem em diferentes contextos (BERBER SARDINHA, 2000).

Desenvolvida com base na pesquisa de Biber sobre a variação de registro na língua inglesa (BIBER, 1988, 1995), a análise multidimensional utiliza a técnica estatística de análise fatorial para reduzir um grande número de características linguísticas a um conjunto menor de dimensões subjacentes. Essas dimensões representam pacotes de características linguísticas que ocorrem juntas e refletem aspectos específicos do estilo ou estrutura da linguagem. Essa abordagem permite a identificação e interpretação de padrões complexos de uso da linguagem que podem ser perdidos em análises mais simples.

Ao longo dos anos, a análise multidimensional tem sido amplamente aplicada em estudos linguísticos, abrangendo diferentes idiomas e áreas de pesquisa, como mudanças ao longo do tempo, linguagem de aprendizes e análise de gêneros. A metodologia desenvolvida

por Biber e seus colegas têm sido fundamental para proporcionar uma compreensão holística e detalhada do uso da linguagem em diferentes contextos e contribui significativamente para o avanço da Linguística de Corpus.

Atualmente, a análise multidimensional é uma metodologia consolidada, usada em diversas áreas de estudos da linguagem, tais como análise do discurso, sociolinguística e linguística aplicada. Conforme os métodos computacionais e tecnologias evoluem, há um promissor potencial de avanço na análise multidimensional e, portanto, a continuidade da relevância e impacto desta análise nos estudos linguísticos.

> "A Análise Multidimensional possui um caráter essencialmente quantitativo e computacional. Ela permite a descrição de línguas e tipos de textos por meio de uma grande quantidade de características linguísticas. [...] A abordagem Multidimensional possui várias características que no seu conjunto distinguem essa metodologia de outros sistemas analíticos de descrição. [...] Ela também tem um caráter essencialmente comparativo, pois promove o contraste entre os textos ou registros. Como diz seu rótulo, ela é multidimensional, ao reconhecer que a variação entre textos e registros pode ser mais adequadamente descrita por meio de múltiplos parâmetros."
>
> (BERBER SARDINHA, 2000)

2.1.5 A consideração do contexto na Linguística de Corpus

O contexto é um aspecto fundamental no estudo da linguagem e desempenha um papel central em todas as áreas que estudam a linguagem. Entender como o contexto influencia a produção e a interpretação de textos e discursos é essencial para compreender a complexidade e a riqueza da comunicação humana.

Na linguística aplicada, o contexto é muitas vezes entendido como o conjunto de fatores sociais, culturais, históricos e situacionais que influenciam e moldam a produção e a interpretação de textos e discursos. Os pesquisadores buscam analisar como diferentes aspectos do contexto influenciam a escolha de palavras, estruturas gramaticais e recursos estilísticos, bem como a interpretação e compreensão de textos por diferentes públicos. Além disso, o contexto é fundamental para a análise do discurso, que examina como a língua é usada em situações comunicativas específicas e como os falantes constroem significados com base em suas experiências, conhecimentos e relações sociais.

A Linguística de Corpus atribui grande importância ao contexto, considerando que "A linguística de corpus pode ser descrita, por enquanto, de forma simples como o estudo da

linguagem com base em exemplos de uso de linguagem "da vida real[3]." (MCENERY & WILSON, 2001, p. 1). Essa importância se dá pela essência da Linguística de Corpus de analisar padrões linguísticos em grandes conjuntos de dados textuais, levando em conta aspectos contextuais que podem influenciar a forma como a língua é usada. Nesse sentido, a Linguística de Corpus adota uma abordagem empírica, baseada em evidências, para estudar a relação entre contexto e linguagem.

Uma das maneiras de levar em conta o contexto na Linguística de Corpus é através da categorização e organização dos dados textuais de acordo com diferentes variáveis contextuais, como gênero discursivo, registro, campo temático, situação comunicativa e perfil sociolinguístico dos falantes. Essa categorização permite que pesquisadores investiguem como diferentes aspectos do contexto podem influenciar a frequência, a distribuição e a combinação de unidades linguísticas, como palavras, colocações, estruturas gramaticais e padrões discursivos.

2.2. Estudos da tradução

A área de estudos da tradução é um campo interdisciplinar que se dedica à análise, pesquisa e prática da tradução, abrangendo tanto a tradução escrita quanto a interpretação oral (BAKER, 1998). Esta área busca entender os processos cognitivos, linguísticos e culturais envolvidos na transferência de significados entre línguas e culturas, bem como as questões éticas, políticas e profissionais relacionadas à prática da tradução. Os estudos da tradução incorporam uma ampla gama de perspectivas teóricas e metodológicas, incluindo abordagens baseadas em linguística, literatura, filosofia, sociologia, psicologia e tecnologia.

Um aspecto importante nos estudos da tradução é a investigação dos diferentes tipos de equivalência que podem ser estabelecidos entre os textos de origem e os de destino. Essa questão aborda a relação entre forma e conteúdo, a adequação das estratégias de tradução e a busca por um equilíbrio entre fidelidade ao texto original e a necessidade de adaptar o texto às características linguísticas, culturais e estilísticas da língua-alvo. Além disso, os estudos da tradução também abordam questões como a visibilidade e o papel do tradutor, a relação entre tradução e poder, e os efeitos da globalização e das tecnologias digitais na prática e na pesquisa da tradução.

A área de estudos da tradução tem crescido rapidamente nas últimas décadas, em parte

[3] Original: *Corpus linguistics is perhaps best described for the moment in simple terms as the study of language based on examples of 'real life' language use*.

devido ao aumento da demanda por serviços de tradução e interpretação em um mundo cada vez mais globalizado e interconectado. Nesse contexto, a formação de tradutores e intérpretes, bem como a pesquisa em tecnologias de tradução automática e ferramentas auxiliares, como os sistemas de tradução assistida por computador (CAT - *Computer Assisted Translation*), têm ganhado maior importância. Ao mesmo tempo, a área de estudos da tradução continua a se aprofundar na compreensão dos aspectos teóricos, práticos e culturais da tradução, buscando contribuir para a qualidade e eficácia da comunicação interlinguística e intercultural.

I. Mapa de Holmes-Toury sobre estudos da tradução

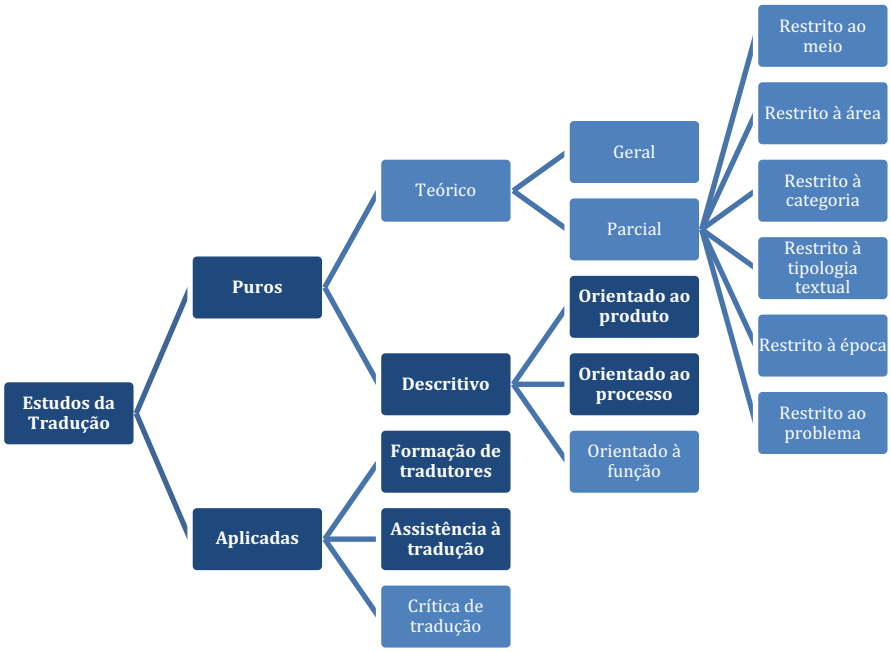
O Mapa de Holmes-Toury sobre estudos da tradução é uma representação visual dos aspectos puros e aplicados dessa área de estudo. Desenvolvido por Gideon Toury (TOURY, 1995) a partir de um estudo seminal de James S. Holmes, o mapa ajudou a moldar a compreensão teórica e prática da disciplina.

No contexto do mapa, os aspectos teóricos e empíricos dos estudos da tradução referem-se à investigação acadêmica da tradução como um fenômeno linguístico, cultural e cognitivo. Isso envolve explorar os processos de transferência de significado entre línguas, a natureza da equivalência entre textos e as questões éticas e políticas relacionadas à tradução. Os aspectos puros visam aprofundar a compreensão dos princípios fundamentais subjacentes à prática da tradução.

Por outro lado, os aspectos aplicados dos estudos da tradução estão relacionados à prática profissional da tradução. Isso inclui o treinamento e desenvolvimento de habilidades de tradução, interpretação e localização, bem como a investigação das implicações práticas da tradução em diferentes contextos. Os aspectos aplicados se concentram em questões como a visibilidade e o papel do tradutor, a utilização de tecnologias de tradução assistida por computador e a relação entre tradução e indústria.

Abaixo, na Figura 2, destacamos no mapa os aspectos mais relevantes desta pesquisa. Em azul escuro estão destacados os aspectos puros e descritivos, tanto orientados ao processo quanto ao produto, e aspectos aplicados, incluindo a formação de tradutores e assistência à tradução.

Figura 2 – Mapa de Holmes-Toury sobre estudos da tradução



Fonte: Adaptado de Baker (1998, p. 278)

Os temas aplicáveis à essa dissertação foram destacados em azul escuro. Em relação aos aspectos puros, a presente pesquisa tem fins descritivos, focados principalmente no produto (traduções humanas e automáticas) como também no processo (modo como tais traduções foram geradas/obtidas). Em relação aos aspectos aplicados, a pesquisa pode servir como base para a formação de novos tradutores assim como ajudar tradutores experientes e iniciantes a aprimorar suas traduções.

2.2.1. Histórico e perspectivas

A prática da tradução remonta a tempos antigos, com registros históricos de traduções entre idiomas como o sumério, o acadiano, o egípcio e o grego. No entanto, a reflexão teórica sobre a tradução começou a se desenvolver durante o período helenístico (323 a.C. - 31 a.C.), quando os tradutores gregos enfrentaram o desafio de traduzir textos hebraicos e aramaicos, como a Septuaginta, para o grego.

Na Idade Média, a tradução adquiriu uma importância central na disseminação do conhecimento e na preservação das tradições clássicas. A tradução de textos religiosos, filosóficos e científicos do grego e do árabe para o latim e outras línguas vernáculas contribuiu para o renascimento europeu e o surgimento de um pensamento humanista.

Os estudos da tradução começaram a se consolidar como uma área de pesquisa acadêmica no século XX, com o desenvolvimento de diversas teorias e abordagens que buscavam compreender a complexidade e a diversidade dos processos e práticas de tradução. Entre os teóricos pioneiros estão Eugene Nida, que propôs os conceitos de equivalência formal e equivalência dinâmica, e Roman Jakobson, que delineou os diferentes tipos de tradução, como a tradução intralingual, interlingual e intersemiótica.

Atualmente, os estudos da tradução são um campo interdisciplinar e em constante evolução, abordando questões que vão desde os aspectos cognitivos e linguísticos da tradução até as dimensões sociopolíticas, culturais e tecnológicas envolvidas no processo. Conforme afirmado por Venuti:

> "*The increasingly interdisciplinary nature of translation studies has multiplied theories of translation. […] In the West, from antiquity to the late nineteenth century, theoretical statements about translation fell into traditionally defined areas of thinking about language and culture […]. Since the beginning of the twentieth century, translation theory has revealed a much expanded range of disciples and approaches in line with the differentiation of modern culture*[5]"
> (VENUTI, 2004, p. 4)

Uma das principais tendências na área de estudos da tradução é o crescente interesse na relação entre tradução e tecnologia. O avanço das tecnologias de tradução automática e das ferramentas de tradução assistida por computador (CAT) tem gerado novas questões e desafios, tanto no que diz respeito à formação de tradutores e intérpretes quanto à ética, qualidade e impacto dessas tecnologias na atuação profissional.

Hutchins é assertivo em sua colocação sobre tradução automática:

> "*The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the late twentieth. [...] Machine translation is not primarily an area of abstract intellectual inquiry but the application of computer and language sciences to the development of systems answering practical needs*[6]."
> (HUTCHINS, 1995, p. 431)

[5] Tradução da autora: A natureza cada vez mais interdisciplinar dos estudos de tradução multiplicou as teorias de tradução. [...] No Ocidente, desde a antiguidade até o final do século XIX, as declarações teóricas sobre a tradução se encaixaram em áreas tradicionalmente definidas de reflexão sobre linguagem e cultura [...]. Desde o início do século XX, a teoria da tradução revelou uma gama muito mais ampla de disciplinas e abordagens, em consonância com a diferenciação da cultura moderna.

[6] Tradução da autora: A tradução de línguas naturais por máquina, sonhada pela primeira vez no século XVII, tornou-se realidade no final do século XX. [...] A tradução automática não é principalmente uma área de investigação intelectual abstrata, mas a aplicação das ciências da computação e da linguagem ao desenvolvimento de sistemas que atendem a necessidades práticas.

Outra tendência importante é a ênfase na interculturalidade e na diversidade linguística. Em um mundo cada vez mais globalizado e interconectado, os estudos da tradução têm se voltado para questões como a tradução de literaturas marginais e minoritárias, a negociação de identidades culturais na tradução e o papel dos tradutores como mediadores interculturais.

Além disso, a área de estudos da tradução explora a crescente importância da localização e da adaptação cultural, especialmente no contexto de produtos e serviços digitais, como videogames, aplicativos e sites. Essa tendência permite que as empresas adaptem seus produtos e serviços para atender às necessidades e expectativas do público-alvo local. Isso inclui a adaptação de elementos linguísticos, culturais e visuais, que podem ser diferentes em diferentes contextos culturais.

A colaboração entre disciplinas também é uma tendência importante nos estudos da tradução. Com a intersecção de áreas como linguística, ciências cognitivas, ciências da computação, estudos culturais e estudos de gênero, a pesquisa em tradução tem se tornado cada vez mais interdisciplinar e abrangente. Essa convergência de perspectivas enriquece o campo e gera novas possibilidades de investigação e aplicação prática.

Em suma, a área de estudos da tradução possui um histórico rico e diversificado, e seu futuro é igualmente promissor. À medida que novas questões e desafios surgem em um mundo cada vez mais interconectado e multilíngue, os estudos da tradução continuarão a evoluir e a se adaptar, buscando fundamentos teóricos e práticos para melhorar a qualidade e a eficácia da comunicação entre línguas e culturas. Essa área de estudo desempenha um papel fundamental na promoção da compreensão intercultural, na disseminação do conhecimento e na construção de pontes entre comunidades e nações, além da constante evolução e transformação do papel do tradutor.

Ao considerar a área de estudos da tradução, é essencial considerar também o papel do tradutor e sua importância ao longo da história, desempenhando uma função importante na comunicação entre culturas e na disseminação de conhecimento. Antigamente, o tradutor geralmente era um erudito que dominava vários idiomas e atuava em diversos campos distintos, incluindo a religião, a filosofia e a literatura. Naquela época, o tradutor era considerado um mediador cultural, permitindo a transferência de conhecimento e ideias entre diferentes culturas e línguas (GENTZLER, 2008).

Traduções mecânicas, nas quais o contexto e cultura são sumariamente ignorados, estão se tornando a norma, pois a qualidade dessas ferramentas está cada vez mais surpreendente. Inicialmente, com a tecnologia, a tradução e o papel do tradutor foram sendo mecanizadas, deixando a mediação cultural em segundo plano. No entanto, quanto mais avança a tecnologia,

menos mecânico se torna o papel do tradutor. Afinal, a máquina está cada vez melhor em produzir traduções mecânicas. Portanto, cabe ao tradutor profissional e qualificado para tal continuamente focar em voltar a exercer o seu papel de mediador cultural, elevando a tradução automática a outro nível.

Atualmente, os tradutores precisam ser proficientes no uso de softwares de tradução, ferramentas de gerenciamento de projetos e outras tecnologias, conforme essas ferramentas evoluem e chegam a níveis cada vez mais próximos ao humano. Os tradutores precisam se manter atualizados com as tendências e as melhores práticas do setor, além de serem capazes de se adaptar rapidamente às mudanças no mercado e às demandas dos clientes. Tudo isso sem perder a essência de uma boa tradução, que inclui qualidade e ética, mas acima de tudo o conhecimento profundo da cultura dos idiomas traduzidos e habilidade primorosa de transferir da melhor forma essa "carga cultural" de um idioma para o outro, seja na tradução técnica ou na tradução literária.

No futuro, é provável que o papel do tradutor continue a evoluir e a mudar com as novas tecnologias e tendências do mercado. À medida que a inteligência artificial e a aprendizagem por máquina continuam a se desenvolver, é possível que as traduções sejam cada vez mais automatizadas, deixando o papel do tradutor para tarefas mais complexas e de alta qualidade. Além disso, a globalização e a crescente demanda por conteúdo multilíngue podem criar novas oportunidades para os tradutores trabalharem em setores e áreas especializadas, como a localização de jogos, a tradução financeira e a tradução de patentes.

Independentemente das mudanças que possam ocorrer, é provável que o papel do tradutor continue essencial para garantir uma comunicação clara e eficaz entre diferentes culturas e idiomas. Os tradutores são fundamentais para garantir que as informações sejam transmitidas de forma precisa e culturalmente apropriada, ajudando a reduzir barreiras linguísticas e a promover a compreensão global. Como tal, é provável que os tradutores continuem sendo uma profissão em demanda no futuro, e aqueles que se adaptam às mudanças e se mantêm atualizados com as tecnologias e tendências emergentes provavelmente terão sucesso em um mercado cada vez mais competitivo.

2.2.2. Tradução e Linguística de Corpus

O histórico da Linguística de Corpus atrelada aos estudos de tradução remonta às últimas décadas do século XX, quando a crescente disponibilidade de computadores e recursos digitais começou a revolucionar a maneira como os linguistas analisavam e trabalhavam com

dados textuais. A combinação dessas duas áreas de estudo tem gerado avanços significativos na pesquisa e na prática da tradução, assim como na formação e no treinamento de tradutores.

A Linguística de Corpus começou a ser aplicada aos estudos de tradução principalmente com o desenvolvimento dos primeiros corpora paralelos. Esses corpora, que consistem em conjuntos de textos originais e suas respectivas traduções, permitem aos pesquisadores comparar diretamente as características linguísticas e estilísticas dos textos em diferentes idiomas e analisar as estratégias de tradução empregadas pelos tradutores.

Um marco importante na aplicação da Linguística de Corpus aos estudos de tradução foi o projeto *Translation Studies Bibliography* (TSB), iniciado na década de 1990 por Mona Baker e outros pesquisadores. Esse projeto desenvolveu um corpus paralelo multilíngue, contendo textos de diversos gêneros e áreas de especialização, o que possibilitou a realização de estudos contrastivos e análises de gêneros textuais.

Outro avanço significativo ocorreu no início dos anos 2000, quando pesquisadores como Silvia Bernardini e Sara Laviosa começaram a aplicar métodos de Linguística de Corpus para investigar padrões e tendências em traduções publicadas (ZANETTIN, BERNARDINI & STEWART, 2003; LAVIOSA, 2002). Nesses estudos, esses e outros autores possibilitaram que os tradutores tivessem acesso a um grande volume de dados linguísticos e culturais, que podem ajudá-los a produzir traduções mais precisas e naturais. Esses estudos revelaram diferenças sistemáticas entre textos traduzidos e não traduzidos, levando ao conceito de "universalidades da tradução" – características comuns que podem ser encontradas em textos traduzidos independentemente das línguas envolvidas.

A crescente interação entre a Linguística de Corpus e os estudos de tradução também tem contribuído para o desenvolvimento de ferramentas e recursos digitais para tradutores e pesquisadores. Por exemplo, a disponibilidade de corpora paralelos e comparáveis facilita o acesso a exemplos de traduções reais e autênticas, permitindo aos tradutores aprimorar suas habilidades e conhecimentos linguísticos.

Além disso, a aplicação da Linguística de Corpus aos estudos de tradução tem sido fundamental para o avanço das tecnologias de tradução automática e tradução assistida por computador (CAT). A análise de grandes conjuntos de dados textuais fornece informações valiosas para treinar algoritmos e desenvolver recursos mais avançados e eficazes para essas ferramentas.

Em resumo, a Linguística de Corpus atrelada aos estudos de tradução tem tido avanços significativos com a crescente interação entre as duas áreas. Essa combinação tem gerado benefícios mútuos, enriquecendo a pesquisa e a prática da tradução, e oferecendo novas

possibilidades e desafios para os pesquisadores, tradutores e educadores envolvidos nesse campo.

À medida que as tecnologias e os recursos digitais continuam a evoluir, é provável que a relação entre a Linguística de Corpus e os estudos de tradução se torne ainda mais estreita e frutífera, abrindo caminho para novas descobertas e inovações. Uma área promissora para a colaboração futura entre a Linguística de Corpus e os estudos de tradução é a integração de métodos quantitativos e qualitativos de análise. A combinação de abordagens estatísticas, provenientes da Linguística de Corpus, com análises interpretativas, típicas dos estudos de tradução, pode proporcionar uma compreensão mais aprofundada dos processos e desafios envolvidos na tradução.

Outro aspecto interessante para a pesquisa futura é a exploração do potencial da Linguística de Corpus para investigar questões relacionadas à ética e à qualidade da tradução. A análise de corpus de textos traduzidos pode fornecer informações importantes sobre como questões como a censura, a manipulação e a adequação cultural são tratadas pelos tradutores em diferentes contextos e épocas.

Além disso, a aplicação da Linguística de Corpus aos estudos de tradução pode oferecer novas perspectivas para a formação e o treinamento de tradutores, ajudando-os a desenvolver habilidades analíticas e críticas para abordar os desafios da tradução no mundo real. A incorporação de técnicas e ferramentas de Linguística de Corpus nos currículos de formação de tradutores pode contribuir para uma abordagem mais orientada para a pesquisa e para o desenvolvimento de competências interdisciplinares, relevantes em diferentes áreas de especialização.

2.2.3. Contexto na área de tradução

A importância do contexto na área de estudos da tradução é inegável e fundamental para o sucesso da tradução. O contexto refere-se a todos os fatores que cercam e influenciam o significado de um texto, incluindo aspectos culturais, sociais, históricos, linguísticos e situacionais.

Compreender e considerar o contexto é essencial para produzir traduções precisas e eficazes, que transmitam adequadamente o significado e a intenção do texto original, de modo a comportar a identificação das nuances culturais e linguísticas que podem afetar o significado de um texto. O contexto cultural é um componente-chave na tradução. Diferentes culturas têm tradições, costumes e sistemas de crenças distintos que influenciam a maneira como as pessoas

se comunicam e interpretam as mensagens. Um bom tradutor deve estar ciente dessas diferenças culturais e conseguir adaptar a tradução de acordo, para garantir que o significado e a intenção do texto original sejam preservados. Isso pode envolver, por exemplo, a adaptação de metáforas, provérbios ou referências culturais específicas, de modo a torná-los compreensíveis e relevantes para o público-alvo (LAVIOSA, 2002).

O contexto social também é crucial na tradução, pois diferentes grupos sociais podem ter maneiras distintas de se expressar e de usar a linguagem. Um tradutor deve considerar o registro linguístico, o tom e o estilo do texto original, bem como o público-alvo e o propósito da tradução. Por exemplo, a tradução de um texto jurídico ou técnico exigirá um registro e um vocabulário específicos, enquanto a tradução de um texto literário ou publicitário pode exigir mais criatividade e flexibilidade na escolha das palavras e expressões (HOUSE, 2006).

Além disso, o contexto histórico e temporal é importante para compreender e traduzir adequadamente um texto. Um tradutor deve estar familiarizado com o período histórico e o ambiente em que o texto original foi produzido, pois isso pode influenciar a linguagem, as referências e os temas abordados. Isso é particularmente relevante ao traduzir textos literários, históricos, audiovisuais ou filosóficos, onde o conhecimento do contexto pode enriquecer a compreensão e a apreciação do texto.

O contexto linguístico também desempenha um papel fundamental na tradução, pois envolve a compreensão das nuances, das ambiguidades e das relações entre as palavras e as expressões no texto original. Um tradutor deve conseguir analisar o texto em termos de sua gramática, sintaxe, semântica e pragmática, e aplicar esse conhecimento para produzir uma tradução precisa e coerente.

O contexto situacional refere-se ao ambiente imediato e à situação em que ocorre a comunicação. Um tradutor deve considerar o propósito e a função do texto original, bem como o contexto em que a tradução será utilizada. Isso pode envolver a adaptação da tradução para atender às necessidades e expectativas do público-alvo ou a consideração de restrições de espaço ou de formato, como na tradução audiovisual ou da localização de software.

Sendo assim, caso o contexto seja ignorado na prática tradutória, isso pode comprometer a qualidade da tradução:

- Perda de significado: A falta de compreensão do contexto pode levar a traduções imprecisas ou incompletas, que não conseguem transmitir adequadamente o significado e a intenção do texto original. Isso pode resultar em mal-entendidos e confusão por parte do público-alvo;

- Ofensa cultural: A desconsideração do contexto cultural pode resultar em traduções culturalmente insensíveis ou ofensivas. Isso pode envolver o uso de expressões inadequadas ou a falta de adaptação de referências culturais específicas, que podem ser mal interpretadas ou mal-recebidas pelos leitores;
- Problemas de estilo e tom: A ausência de contexto pode levar a traduções que não respeitam o estilo, o tom e o registro linguístico do texto original, o que pode prejudicar a coerência e a fluidez da tradução. Isso é particularmente relevante ao traduzir textos literários, poéticos, audiovisuais ou publicitários, onde o estilo e o tom são aspectos cruciais da comunicação;
- Perda de nuances e ambiguidades: Sem considerar o contexto, os tradutores podem perder as nuances e ambiguidades presentes no texto original, levando a uma tradução que não capta a riqueza e a complexidade do texto-fonte. Isso pode ser especialmente problemático na tradução de textos literários, humorísticos ou poéticos, onde as sutilezas da linguagem desempenham um papel importante;
- Falta de adaptação ao público-alvo: A ausência de contexto pode resultar em traduções que não são adaptadas às necessidades e expectativas do público-alvo. Isso pode envolver a falta de consideração de diferenças culturais, sociais ou de registro linguístico, que podem afetar a relevância e a aceitabilidade da tradução;
- Erros históricos ou factuais: Ignorar o contexto histórico ou temporal pode levar a traduções que contêm erros factuais ou que não refletem adequadamente o ambiente em que o texto original foi produzido. Isso pode ser especialmente prejudicial ao traduzir textos históricos, acadêmicos ou jornalísticos, onde a precisão e a veracidade são essenciais;
- Comprometimento da qualidade da tradução: A falta de contexto pode, em última análise, comprometer a qualidade geral da tradução, tornando-a menos eficaz, coerente e atraente para os leitores. Isso pode ter implicações negativas para a reputação e a carreira dos tradutores, bem como para a percepção e o entendimento entre diferentes culturas e comunidades.

De acordo com Fukari & Wolf:

> "*Norms operate in each phase of the Translation process [...]. A detailed analysis of all translation norms effective at a specific time within a specific society would ideally enable insights into that society's ideas on translation as a cultural phenomenon. […] The 'agreements and conventions' underlying the practice of translation are continuously negotiated by the people and institutions involved*[7]."
> (FUKARI & WOLF, 2007, p. 9)

Ao considerar os diferentes aspectos do contexto e aplicar esse conhecimento ao longo do processo de tradução, os tradutores podem produzir traduções que não apenas coloquem em jogo o significado e a intenção do texto fonte, mas também respeitam e refletem as especificidades culturais, sociais e situacionais do público-alvo.

2.2.4. Avanços tecnológicos na área de tradução

Os avanços tecnológicos na área de tradução têm impulsionado significativas mudanças nos estudos de tradução, impactando a maneira como os tradutores trabalham e como os pesquisadores abordam a análise e a compreensão do processo de tradução. Uma dessas mudanças é a introdução de ferramentas de Tradução Assistida por Computador (CAT), como os sistemas de gerenciamento de terminologia e as memórias de tradução.

Uma mudança de paradigma importante nos estudos de tradução é o advento e a popularização da tradução automática (MT), especialmente com o desenvolvimento de sistemas baseados em Inteligência Artificial (IA) e Aprendizagem por Máquina (ML). Essas tecnologias levam a melhorias significativas na qualidade da tradução automática, permitindo traduções mais rápidas e, em alguns casos, mais precisas.

Isso tem gerado novos desafios e oportunidades para os tradutores e pesquisadores, que agora precisam investigar a interação entre tradução humana e automática e desenvolver estratégias para garantir a qualidade e a eficácia das traduções produzidas. As repercussões dessas transformações também afetam a maneira como as empresas e organizações lidam com

[7] Tradução da autora: As normas operam em cada fase do processo de tradução [...]. Uma análise detalhada de todas as normas de tradução efetivas em um momento específico dentro de uma sociedade específica idealmente permitiria insights sobre as ideias dessa sociedade em relação à tradução como fenômeno cultural. [...] Os 'acordos e convenções' subjacentes à prática da tradução são continuamente negociados pelas pessoas e instituições envolvidas.

a tradução de documentos e informações, permitindo uma maior eficiência e produtividade (WAY, 2018).

Por fim, a crescente interação entre Linguística de Corpus e estudos de tradução tem proporcionado novos métodos e abordagens para analisar e entender o processo de tradução. A disponibilidade de grandes volumes de dados textuais e o desenvolvimento de ferramentas computacionais para analisá-los possibilitaram aos pesquisadores investigar padrões linguísticos e culturais em diferentes idiomas e textos traduzidos. Essa abordagem baseada em dados enriquece os estudos de tradução, permitindo uma compreensão mais aprofundada e abrangente das complexidades e desafios envolvidos na tradução em um mundo cada vez mais globalizado e interconectado (BAKER, 1995).

Abaixo detalhamos o histórico, perspectivas e aplicabilidades dos avanços tecnológicos na área de tradução.

2.2.4.1. Tradução assistida por computador

A história da tradução assistida por computador (CAT - *Computer-Aided Translation*) remonta às últimas décadas do século XX, quando os avanços tecnológicos possibilitaram o desenvolvimento de ferramentas de software projetadas para auxiliar os tradutores no processo de tradução (O'HAGAN & ASHWORTH, 2002; PIETRZAK & KORNACKI, 2020). As CAT *tools* visam a melhorar a eficiência e a qualidade das traduções, fornecendo recursos como memórias de tradução, glossários e verificações automáticas de consistência.

Uma das primeiras CAT *tools* foi o programa ALPS (*Automated Language Processing Systems*), desenvolvido na década de 1980. Este software pioneiro oferecia funções como processamento de texto e recursos de dicionário eletrônico, permitindo que os tradutores consultassem rapidamente termos e expressões em diferentes idiomas. O ALPS foi seguido por outras ferramentas semelhantes, como o *Translator's Workbench*, lançado pela Trados em 1987, que introduziu o conceito de memória de tradução.

As memórias de tradução se tornaram uma característica central das CAT *tools*, pois permitem que os tradutores armazenem e reutilizem segmentos de texto traduzidos anteriormente. Essa tecnologia ajuda a garantir a consistência e a qualidade das traduções e permite aos tradutores economizar tempo e esforço, especialmente ao trabalhar com documentos repetitivos ou altamente técnicos.

Ao longo dos anos 1990 e 2000, as CAT *tools* continuaram a evoluir e a se diversificar. Surgiram várias ferramentas comerciais, como o Wordfast e o Trados, cada uma com suas

próprias características e funcionalidades. Além disso, foram desenvolvidas ferramentas de código aberto, como o OmegaT, que permitiam maior personalização e acesso a um público maior de tradutores.

Outro desenvolvimento importante na história das CAT tools é a integração com outras tecnologias e recursos, como a tradução automática e a localização de software e websites. Muitas ferramentas modernas de CAT incluem opções para pré-traduzir segmentos de texto usando tradução automática, que os tradutores podem revisar e editar conforme necessário. Considerando esses fatores, as CAT *tools* são uma parte essencial da indústria da tradução moderna, permitindo a produção de traduções mais eficientes e precisas em diferentes línguas e culturas.

Em resumo, a história das CAT *tools* é marcada por inovação contínua e adaptação às necessidades dos tradutores e do mercado em constante mudança. Hoje, as CAT tools são uma parte essencial do trabalho de muitos tradutores profissionais, permitindo que lidem com projetos de tradução complexos e demorados de maneira mais eficiente e eficaz. Com a evolução constante da tecnologia, as CAT *tools* provavelmente continuarão a se desenvolver e a se adaptar às demandas do setor de tradução no futuro.

As perspectivas futuras para as CAT *tools* são promissoras, à medida que a tecnologia avança e o setor de tradução continua a crescer e se adaptar às necessidades globais. Um dos principais focos para o futuro das CAT *tools* é a integração ainda maior com a tradução automática e a inteligência artificial, proporcionando um ambiente de tradução mais fluido e eficiente, onde os tradutores podem se concentrar no aprimoramento e na localização de conteúdo gerado automaticamente.

Além disso, é provável que as CAT *tools* se tornem ainda mais personalizáveis e adaptáveis, permitindo que os tradutores configurem suas próprias soluções e fluxos de trabalho de acordo com suas necessidades específicas. Isso pode incluir a integração com outras ferramentas e aplicativos, bem como a utilização de plug-ins e extensões para adicionar funcionalidades específicas ou melhorar a interoperabilidade entre diferentes plataformas e formatos de arquivo.

Há também perspectivas otimistas em relação aos arquivos que as CAT *tools* serão capazes de processar no futuro e também suas funcionalidades, como diferentes tipos de dados, como áudio e vídeo, permitindo a produção de traduções multimodais e uma experiência mais imersiva para o usuário final.

No que diz respeito aos desafios enfrentados pelas CAT *tools*, um dos principais problemas é garantir a qualidade e a precisão das traduções geradas ou processadas por essas

ferramentas. Embora a tecnologia de tradução automática tenha melhorado significativamente nos últimos anos, ainda há limitações na capacidade das máquinas de compreender e reproduzir nuances culturais e linguísticas complexas. Portanto, é essencial que as CAT *tools* continuem a evoluir para oferecer aos tradutores recursos e funcionalidades que os ajudem a garantir a qualidade e a precisão das traduções (BOWKER, 2002).

Outro desafio é garantir que as CAT *tools* sejam acessíveis para tradutores com diferentes níveis de experiência e conhecimento tecnológico. Isso pode incluir o desenvolvimento de interfaces de usuário mais intuitivas e a disponibilização de materiais de treinamento e suporte abrangentes para ajudar os tradutores a aproveitar ao máximo as funcionalidades oferecidas por essas ferramentas.

As aplicabilidades e vantagens das CAT *tools* para tradutores e o público são diversas. Em primeiro lugar, as CAT *tools* melhoram significativamente a eficiência dos tradutores, permitindo-lhes processar grandes volumes de texto em menos tempo. A reutilização de segmentos de texto traduzidos anteriormente por meio de memórias de tradução ajuda a garantir a consistência e a qualidade das traduções, além de economizar tempo e esforço.

As ferramentas de tradução assistida por computador são projetadas para ajudar os tradutores humanos a trabalhar de forma mais rápida e eficiente, proporcionando uma série de recursos e funcionalidades úteis. Algumas das principais funcionalidades das CAT tools incluem:

- Memórias de tradução (TM): As memórias de tradução armazenam segmentos de texto previamente traduzidos e os tornam disponíveis para reutilização em traduções futuras. Isso permite que os tradutores economizem tempo e garantam a consistência em projetos de tradução, especialmente em documentos com conteúdo repetitivo. Além disso, é possível utilizar as memórias de tradução como base de pesquisa terminológica e como base para novas traduções.
- Glossários e terminologia: As CAT tools permitem que os tradutores criem e gerenciem glossários de termos específicos do setor ou da empresa, garantindo o uso correto e consistente da terminologia em todas as traduções.
- Verificação de consistência: As CAT tools podem verificar automaticamente a consistência do uso de termos e formatação em um

documento traduzido, ajudando a identificar e corrigir possíveis erros ou inconsistências.

- Segmentação de texto: As CAT tools segmentam automaticamente o texto-fonte em partes menores, como frases ou parágrafos, facilitando a tradução e a revisão.
- Alinhamento de texto: As CAT tools podem alinhar automaticamente o texto-fonte e o texto traduzido, facilitando a comparação entre os dois e a identificação de possíveis erros ou melhorias na tradução.
- Suporte a formatos de arquivo: As CAT tools são compatíveis com uma ampla variedade de formatos de arquivo, incluindo documentos de texto, planilhas, apresentações e conteúdo de websites, permitindo que os tradutores trabalhem com diferentes tipos de mídia e garantam a consistência das traduções em várias plataformas.
- Localização: As CAT tools fornecem recursos específicos para ajudar os tradutores a adaptar produtos e serviços a mercados e públicos-alvo específicos, levando em consideração aspectos culturais e linguísticos, como unidades de medida, formatos de data e hora e moedas.
- Gerenciamento de projetos: As CAT tools também oferecem recursos de gerenciamento de projetos, permitindo que os tradutores e gerentes de projetos monitorem o progresso das traduções, atribuam tarefas a outros tradutores e garantam o cumprimento dos prazos e requisitos de qualidade.
- Colaboração em tempo real: Algumas CAT tools avançadas oferecem recursos de colaboração em tempo real, permitindo que vários tradutores trabalhem simultaneamente em um mesmo projeto, compartilhando memórias de tradução, glossários e outras informações relevantes.

Para o público, as CAT *tools* têm o potencial de aumentar a disponibilidade e a acessibilidade de conteúdo traduzido em diversos idiomas. À medida que a demanda por traduções de alta qualidade cresce, as CAT tools permitem que os tradutores atendam a essa demanda de maneira mais eficiente, tornando mais fácil para as pessoas acessarem informações, conhecimentos e recursos culturais em idiomas diferentes do seu próprio.

As CAT *tools* podem desempenhar um papel crucial na localização de produtos e serviços, ajudando empresas e organizações a se adaptarem às necessidades e preferências específicas de diferentes mercados e públicos-alvo. Ao fornecer recursos para gerenciar e

adaptar aspectos como formatos de data e hora, unidades de medida e moeda, as CAT *tools* ajudam a garantir que as traduções sejam culturalmente apropriadas e facilmente compreensíveis para os usuários finais.

A integração das CAT *tools* com a tradução automática também oferece oportunidades para a colaboração entre tradutores humanos e máquinas. Ao combinar a velocidade e a eficiência da tradução automática com a habilidade e o conhecimento cultural dos tradutores humanos, é possível alcançar resultados mais rápidos e precisos. Isso pode ser particularmente útil em situações em que o tempo é essencial, como na tradução de notícias ou documentos de emergência.

Em suma, as CAT *tools* proporcionam um compêndio de vantagens e aplicabilidades tanto para profissionais da tradução quanto para o público em geral, contribuindo para a eficiência, a qualidade e a acessibilidade das traduções em um contexto mundial cada vez mais interligado e globalizado. Com o avanço contínuo da tecnologia e a crescente demanda por traduções em diversos setores, é provável que as CAT *tools* continuem a desempenhar um papel cada vez mais importante na facilitação da comunicação e do entendimento recíproco entre distintas culturas e línguas.

2.2.4.2 Tradução automática

A história da tradução automática (TA) moderna começa no século XX, com os primeiros esforços teóricos para desenvolver métodos e algoritmos capazes de traduzir automaticamente textos entre diferentes idiomas. O interesse em desenvolver sistemas de TA foi intensificado pela necessidade de comunicação eficiente durante e após a Segunda Guerra Mundial, quando a troca de informações entre nações e a necessidade de processar grandes volumes de documentos em diferentes idiomas tornaram-se cruciais.

A Guerra Fria foi um período particularmente importante para o desenvolvimento da tradução automática. Durante esse período, a corrida armamentista e a competição tecnológica entre os Estados Unidos e a União Soviética levaram a um aumento significativo nos investimentos em pesquisa e desenvolvimento na área de processamento de linguagem natural (PLN) e tradução automática.

Nessa época, a tradução automática foi usada pela Agência de Segurança Nacional dos EUA durante a Guerra Fria para monitorar comunicações em diferentes línguas e culturas, permitindo uma maior capacidade de detecção de ameaças inimigas (HUTCHINS & SOMERS,

1992). A ideia era que a tradução automática também pudesse ajudar a analisar a vasta quantidade de informações produzidas pelo "inimigo" em um idioma desconhecido. Além disso, a corrida espacial também impulsionou a pesquisa em TA, já que a comunicação entre astronautas e cientistas de diferentes países exigia uma tradução rápida e precisa.

Os primeiros sistemas de tradução automática, desenvolvidos nas décadas de 1950 e 1960, baseavam-se em abordagens baseadas em regras e dicionários bilíngues, que tentavam mapear estruturas gramaticais e léxicas entre os idiomas. Esses sistemas eram limitados em sua capacidade de lidar com ambiguidades e nuances linguísticas e, muitas vezes, produziam traduções de baixa qualidade. No entanto, eles estabeleceram as bases para o desenvolvimento de abordagens mais sofisticadas e eficientes no futuro.

A partir dos anos 1980 e 1990, a pesquisa em tradução automática começou a se concentrar em abordagens baseadas em estatísticas e exemplos, que utilizavam grandes corpora de textos bilíngues para treinar modelos probabilísticos de tradução. Esses métodos permitiram avanços significativos na qualidade das traduções automáticas, embora ainda enfrentassem desafios na captura de nuances e ambiguidades linguísticas.

O maior salto na qualidade da tradução automática ocorreu na última década, com o advento das abordagens baseadas em aprendizado profundo e redes neurais. Os sistemas de tradução automática neural, como o *Google Neural Machine Translation* (GNMT), um exemplo de como a tecnologia de inteligência artificial está transformando a indústria da tradução, permitiu a produção de traduções mais precisas e contextuais em diferentes línguas e culturas (WU et. al., 2016).

Os recursos de tradução automática conseguem aprender representações semânticas e sintáticas complexas a partir de grandes volumes de dados e gerar traduções de alta qualidade em tempo real. Essas tecnologias revolucionam a indústria da tradução e continuam a impulsionar a pesquisa e o desenvolvimento na área de tradução automática.

As perspectivas para a tradução automática (TA) são promissoras, com avanços tecnológicos contínuos e crescente interesse no desenvolvimento e aplicação dessa tecnologia em diversos campos. No entanto, a TA também enfrenta desafios e questões relacionadas à qualidade, à ética e ao impacto no mercado de trabalho. Neste contexto, é fundamental explorar as oportunidades, os desafios e as aplicações potenciais da TA.

As perspectivas futuras para a tradução automática incluem aprimoramentos na qualidade das traduções, maior abrangência de idiomas, especialmente para idiomas menos difundidos ou com poucos recursos, e integração aprofundada com outras tecnologias e campos de estudo, como a inteligência artificial, o aprendizado por máquina e a análise de sentimentos.

Além disso, espera-se que a TA desempenhe um papel fundamental na promoção da comunicação e compreensão intercultural em um mundo cada vez mais globalizado.

No entanto, os desafios persistem. A qualidade das traduções automáticas, apesar dos avanços recentes, ainda enfrenta dificuldades em lidar com ambiguidades, nuances culturais e especificidades linguísticas. Além disso, questões éticas relacionadas à privacidade, à segurança e ao uso indevido da tecnologia de TA também precisam ser abordadas e gerenciadas cuidadosamente.

Outro desafio é o impacto da TA no mercado de trabalho para tradutores humanos. Embora a TA possa aumentar a eficiência e a produtividade dos tradutores, também existe o risco de que alguns trabalhos sejam substituídos por máquinas, especialmente em tarefas mais rotineiras e simples. Portanto, é crucial que os tradutores se adaptem às novas tecnologias e desenvolvam habilidades complementares para se manterem competitivos no mercado.

As aplicações da tradução automática são vastas e variadas, abrangendo diversos setores e contextos, tornando-se uma ferramenta valiosa na era da globalização e da comunicação digital. As possibilidades de uso da TA continuam a evoluir conforme a tecnologia avança e se torna mais sofisticada.

No âmbito empresarial, a TA desempenha um papel fundamental ao permitir que empresas e organizações se expandam para mercados internacionais, traduzindo documentos, contratos, manuais, materiais de marketing e comunicações entre funcionários e clientes de diferentes origens linguísticas. Isso facilita a colaboração, aumenta a produtividade e promove o crescimento dos negócios em escala global.

Na área da educação e do aprendizado de idiomas, a tradução automática pode ser uma ferramenta útil para estudantes e professores, fornecendo traduções rápidas e, em muitos casos, precisas de materiais educacionais e recursos. Além disso, a TA pode ajudar os alunos a adquirir vocabulário e compreensão em outros idiomas, permitindo que acessem informações e conhecimentos em idiomas diferentes do seu próprio. Isso pode ser especialmente útil em cursos e programas de estudo que abordam tópicos globais e interculturais.

Outra aplicabilidade importante da tradução automática é na área de acesso à informação e ao conhecimento. Com a TA, torna-se mais fácil para as pessoas acessarem informações e conhecimentos em outros idiomas, ampliando o alcance da produção científica, cultural e literária. Isso pode incluir a tradução de artigos acadêmicos, notícias, conteúdo online e obras literárias, permitindo que um público mais amplo se beneficie desses recursos e contribua para a disseminação do conhecimento e do entendimento intercultural.

A TA também pode ser utilizada no contexto das redes sociais e da comunicação online, ajudando indivíduos e comunidades a se conectarem e a compartilharem ideias, experiências e informações através das barreiras linguísticas. Plataformas de mídia social e aplicativos de mensagens frequentemente incorporam recursos de tradução automática, permitindo que os usuários se comuniquem com pessoas de diferentes origens linguísticas e culturas, promovendo a compreensão e a cooperação internacionais. Em resumo, a tradução automática tem um enorme potencial de aplicabilidade em várias áreas da vida moderna, melhorando a comunicação e o entendimento entre pessoas de todo o mundo.

Em conclusão, a tradução automática tem se mostrado uma ferramenta poderosa e versátil em uma variedade de setores, desde negócios e educação até comunicação online e acesso à informação. A TA tem o potencial de transformar a maneira como nos comunicamos e interagimos globalmente, promovendo a compreensão intercultural e a cooperação. No entanto, é importante abordar os desafios e questões éticas relacionadas à qualidade e ao impacto no mercado de trabalho, buscando aprimorar a tecnologia e explorar novas aplicações e oportunidades. Ao fazer isso, a tradução automática pode continuar a contribuir significativamente para um mundo cada vez mais conectado e diversificado.

2.2.4.3. Pós-edição de tradução automática

A pós-edição de tradução automática é um campo que se desenvolveu ao longo das últimas décadas, acompanhando os avanços na área da tradução automática (TA). Os primórdios da pós-edição remontam à década de 1950, quando as primeiras tentativas de tradução automática começaram a surgir. Essas primeiras abordagens, baseadas em regras gramaticais e léxicas, mostraram-se limitadas e, como resultado, os tradutores humanos eram frequentemente necessários para revisar e corrigir as traduções geradas por máquinas (O'BRIEN, 2011).

A pós-edição ganhou maior relevância nas décadas de 1980 e 1990, com o desenvolvimento de sistemas de TA baseados em estatísticas. Esses sistemas, que utilizam algoritmos de aprendizagem por máquina, melhoraram a qualidade das traduções automáticas, mas ainda exigiam a intervenção humana para garantir a precisão e a fluência dos textos traduzidos

O termo "pós-edição" foi oficialmente reconhecido em 1992, quando a Associação Europeia de Tradução Automática (EAMT - *European Association for Machine Translation*) definiu-o como o processo de melhorar e corrigir as traduções automáticas por um tradutor

humano. A partir de então, a pós-edição começou a se consolidar como uma disciplina específica dentro do campo da tradução e da localização. Na década de 2000, a pós-edição passou por uma rápida evolução, impulsionada pela popularização da internet e pela crescente demanda por conteúdo multilíngue. Essa demanda levou à criação de novas ferramentas e tecnologias de tradução automática, como o Google Translate, que foram amplamente adotadas por usuários e empresas em todo o mundo.

A pós-edição de tradução automática pode ser vista como uma oportunidade para os tradutores, permitindo que eles usem suas habilidades linguísticas e culturais para garantir a qualidade e fluência da tradução final, ao mesmo tempo em que usam a tecnologia de tradução automática como uma ferramenta de suporte.

Em 2013, a norma europeia de qualidade para serviços de tradução (EN 15038) foi substituída pela norma internacional ISO 17100, que estabeleceu diretrizes específicas para a pós-edição de tradução automática. A norma ISO 18587, publicada em 2017, forneceu ainda mais orientações sobre os requisitos e as melhores práticas para a pós-edição de TA.

Embora a ISO 17100 e a ISO 18587 sejam duas normas diferentes, ambas têm um papel importante na indústria da tradução e na área de pós-edição de tradução automática. Abaixo, detalhamos as principais diretrizes de cada uma dessas normas.

I. ISO 17100:2015 - Serviços de tradução - Requisitos para os serviços de tradução

A ISO 17100 (BRITISH STANDARDS INSTITUTION, 2015) é uma norma internacional que estabelece os requisitos para os serviços de tradução. Embora não se concentre especificamente na pós-edição, ela estabelece um quadro geral para garantir a qualidade dos serviços de tradução, incluindo a tradução automática. Alguns dos principais aspectos da ISO 17100 são:

- Requisitos para os recursos humanos: a norma específica a competência dos tradutores, revisores e outros profissionais envolvidos no processo de tradução, incluindo a "formação, qualificação e experiência profissional, destacando a importância de garantir a competência dos tradutores [...] e recomenda que os tradutores sejam treinados e capacitados para trabalhar com tecnologias de tradução automática." (ISO 17100:2015, p. 6)

- Processos de tradução: a ISO 17100 estabelece diretrizes para o processo de tradução em si, que inclui a “preparação, a tradução, a revisão, a verificação e a entrega do projeto, estabelecendo, por exemplo, que caso a tradução automática seja utilizada, ela seja antes avaliada em relação à sua adequação para o propósito, antes de ser usada em qualquer projeto de tradução." (ISO 17100:2015, p. 6)
- Gestão da qualidade: a norma ISO 17100 enfatiza a importância da gestão da qualidade e do monitoramento contínuo do desempenho dos serviços de tradução, estabelecendo que a tradução automática deve ser pós-editada por um tradutor humano qualificado para garantir a qualidade e a precisão da tradução final.

II. ISO 18587:2017 - Serviços de tradução - Pós-edição de tradução automática

A ISO 18587 (BRITISH STANDARDS INSTITUTION, 2017) é uma norma específica para a pós-edição de tradução automática e estabelece “requisitos claros para a documentação e comunicação entre os provedores de serviços de tradução, os pós-editores de tradução automática e os clientes, garantindo a transparência e a rastreabilidade do processo de pós-edição de tradução automática". (ISO 18587:2017, p. 1). Algumas das principais diretrizes da ISO 18587 incluem:

- Definição de pós-edição: a norma define pós-edição como o processo de revisão e correção de traduções automáticas por um pós-editor humano.
- Requisitos para pós-editores: a norma estabelece os requisitos de competência para os pós-editores, que incluem habilidades linguísticas, conhecimento cultural, habilidades técnicas e conhecimento do assunto.
- Processo de pós-edição: a ISO 18587 descreve o processo de pós-edição, incluindo a preparação, a avaliação da qualidade da tradução automática, a pós-edição em si e a verificação da qualidade do texto pós-editado.
- Comunicação entre partes envolvidas: a norma enfatiza a importância da comunicação clara e eficaz entre o cliente, o fornecedor de serviços de tradução e o pós-editor.

- Garantia de qualidade e monitoramento: a ISO 18587 destaca a importância de garantir e monitorar a qualidade da pós-edição e implementar melhorias no processo conforme necessário.

Em resumo, a ISO 17100 estabelece um quadro geral para os serviços de tradução, enquanto a ISO 18587 se concentra especificamente na pós-edição de tradução automática. Ambas as normas são fundamentais para garantir a qualidade e a eficiência dos serviços de tradução e pós-edição.

A introdução de sistemas de tradução automática baseados em redes neurais, a partir de 2016, marcou uma nova era na pós-edição. Esses sistemas, conhecidos como tradução automática neural (NMT), geram traduções de maior qualidade e mais naturais, o que reduz o esforço necessário para a pós-edição.

A crescente aceitação e adoção da pós-edição de TA levaram à formação de uma indústria específica, com empresas especializadas na oferta de serviços de pós-edição. Além disso, muitas universidades e instituições de ensino passaram a oferecer programas e cursos especializados em pós-edição de tradução automática.

A indústria de pós-edição continua a enfrentar desafios, como a necessidade de desenvolver melhores ferramentas e tecnologias de apoio, aprimorar a formação de profissionais e aperfeiçoar os processos de trabalho. Um dos principais desafios é a integração eficiente entre os sistemas de TA e as ferramentas de pós-edição. Para enfrentar esses desafios, a comunidade de pós-edição trabalha em estreita colaboração com pesquisadores e desenvolvedores na busca por soluções inovadoras e novas abordagens para melhorar a qualidade e a eficiência do processo de pós-edição.

À medida que a tradução automática continua a avançar, a pós-edição se consolida como uma área de estudo e prática essencial para garantir que as traduções automáticas atendam aos padrões de qualidade exigidos pelos usuários e pelos setores profissionais. O futuro da pós-edição provavelmente envolverá a adoção de técnicas avançadas de inteligência artificial e aprendizagem por máquina, assim como a maior integração entre os profissionais de pós-edição e os sistemas de TA. Essa evolução constante garantirá que a pós-edição de tradução automática continue desempenhando um papel vital na comunicação global e na produção de conteúdo multilíngue de alta qualidade.

2.2.5 Tradução financeira

A tradução técnica na área financeira é uma área altamente especializada que requer o uso de ferramentas de tradução e recursos especializados, como dicionários financeiros e glossários terminológicos, para garantir a precisão e a qualidade da tradução final, permitindo que empresas e instituições financeiras se comuniquem de forma eficiente e precisa em diferentes idiomas. Essa área de tradução envolve a conversão de documentos, relatórios e materiais relacionados às finanças e à economia, garantindo que a terminologia, os conceitos e a precisão numérica sejam preservados no idioma-alvo.

A tradução financeira é um campo que exige habilidades de pesquisa e consulta a fontes especializadas para garantir a precisão e a exatidão da tradução. Os tradutores especializados nessa área devem possuir não apenas habilidades linguísticas avançadas, mas também um sólido conhecimento dos conceitos financeiros, das regulamentações locais e internacionais e da terminologia específica do setor. Além disso, é importante que o tradutor esteja atualizado em relação às tendências e às mudanças no cenário financeiro global, já que isso afeta diretamente a precisão e a relevância das traduções.

A precisão nessa área é de suma importância. Erros ou ambiguidades podem levar a interpretações incorretas, o que pode resultar em perdas financeiras significativas ou problemas legais. Portanto, os tradutores financeiros devem ser extremamente meticulosos e detalhistas ao traduzir números, taxas, prazos e termos técnicos. Além disso, é fundamental garantir a consistência terminológica em todos os documentos e materiais traduzidos.

Para empresas de capital aberto, a tradução adquire uma importância ainda maior, uma vez que essas companhias são obrigadas a divulgar informações financeiras detalhadas e precisas aos acionistas, investidores e órgãos reguladores. A tradução de documentos corporativos deve ser realizada com extrema precisão e clareza para garantir a transparência e a conformidade com as leis e regulamentações aplicáveis em cada jurisdição.

Uma das particularidades da tradução financeira para empresas de capital aberto é a necessidade de se adaptar às normas contábeis e regulatórias específicas dos países em que a empresa opera. Dependendo do mercado, as empresas podem estar sujeitas a diferentes padrões contábeis, como o *International Financial Reporting Standards* (IFRS) ou o *Generally Accepted Accounting Principles* (GAAP) dos Estados Unidos. Os tradutores financeiros devem estar cientes dessas diferenças e garantir que os documentos traduzidos reflitam corretamente os padrões contábeis e regulatórios aplicáveis.

Além disso, as empresas de capital aberto frequentemente realizam fusões e aquisições, o que envolve a tradução de um grande volume de documentos financeiros e legais. Nesses casos, a tradução desempenha um papel crucial na facilitação da comunicação entre as partes

envolvidas e na garantia de que os termos e condições das transações sejam compreendidos de maneira clara e inequívoca.

A tradução também desempenha um papel fundamental para as empresas de capital aberto em relação à necessidade de sincronização na divulgação de informações em todos os mercados nos quais a empresa está listada. Essa sincronização é crucial para garantir que todos os investidores, acionistas e órgãos reguladores em diferentes países e regiões tenham acesso às mesmas informações financeiras e corporativas ao mesmo tempo, promovendo a igualdade de condições e a transparência no mercado global.

Quando uma empresa de capital aberto divulga informações, como resultados financeiros, fusões e aquisições, mudanças na gestão ou outros anúncios relevantes, é importante que essas informações sejam comunicadas de maneira clara, precisa e simultânea a todos os interessados, independentemente do idioma e da localização. Isso garante que os investidores possam tomar decisões informadas e minimiza o risco de desinformação ou desigualdade de acesso às informações, que poderiam levar a distorções no mercado e à especulação.

A tradução de alta qualidade é essencial para alcançar essa sincronização, pois garante que os documentos financeiros e corporativos sejam compreendidos corretamente pelos públicos-alvo em diferentes países e regiões. Tradutores especializados em finanças devem conseguir adaptar adequadamente a linguagem, os conceitos e as práticas contábeis às especificidades culturais e regulatórias de cada mercado, sem comprometer a precisão e a clareza das informações.

Além disso, a tradução eficiente e a sincronização na divulgação de informações podem ajudar as empresas de capital aberto a cumprir com as leis e regulamentações locais e internacionais, como a diretiva de transparência da União Europeia ou as regras da *Securities and Exchange Commission* (SEC) nos Estados Unidos ou da Comissão de Valores Mobiliários (CVM) no Brasil. Essas regulamentações frequentemente exigem que as empresas divulguem informações financeiras e corporativas de maneira oportuna e transparente em todos os mercados em que operam.

Um exemplo notável de como uma tradução incorreta pode gerar problemas graves para uma empresa de capital aberto ocorreu em 2011 com a fabricante japonesa de automóveis Honda. A empresa enfrentou uma crise de relações públicas e perdas financeiras significativas devido a um erro de tradução em uma de suas apresentações financeiras.

Na época, a Honda estava se recuperando de um desastre natural que afetou gravemente sua capacidade de produção. A empresa realizou uma conferência de imprensa para apresentar

seus resultados financeiros e discutir as perspectivas futuras. Durante a apresentação, um executivo da Honda mencionou que a empresa estava trabalhando para "retomar a produção total" em suas fábricas. No entanto, a declaração foi traduzida incorretamente para o inglês como "voltar à produção normal", o que levou os investidores e a mídia a acreditar que a Honda estava prestes a retomar sua produção ao nível anterior ao desastre.

Essa tradução incorreta gerou uma onda de otimismo em relação à recuperação da Honda, fazendo com que as ações da empresa subissem significativamente. No entanto, quando a empresa esclareceu posteriormente a declaração e confirmou que a produção ainda estava longe de voltar ao normal, as ações caíram abruptamente, resultando em perdas financeiras significativas para os acionistas e danos à reputação da empresa.

Em conclusão, a tradução financeira é um aspecto crucial na comunicação eficiente e precisa entre empresas, investidores e órgãos reguladores em um mundo globalizado. Especialmente no caso de empresas de capital aberto, traduções precisas são essenciais para garantir a transparência, a conformidade com as regulamentações e a tomada de decisões informadas. Erros de tradução podem levar a consequências financeiras e de reputação significativas, como ilustrado no exemplo da Honda. Portanto, é de suma importância que as empresas invistam em tradutores especializados e qualificados na área financeira, bem como em processos de garantia de qualidade, como revisões e verificações, para assegurar a clareza e a precisão em todas as suas comunicações financeiras.

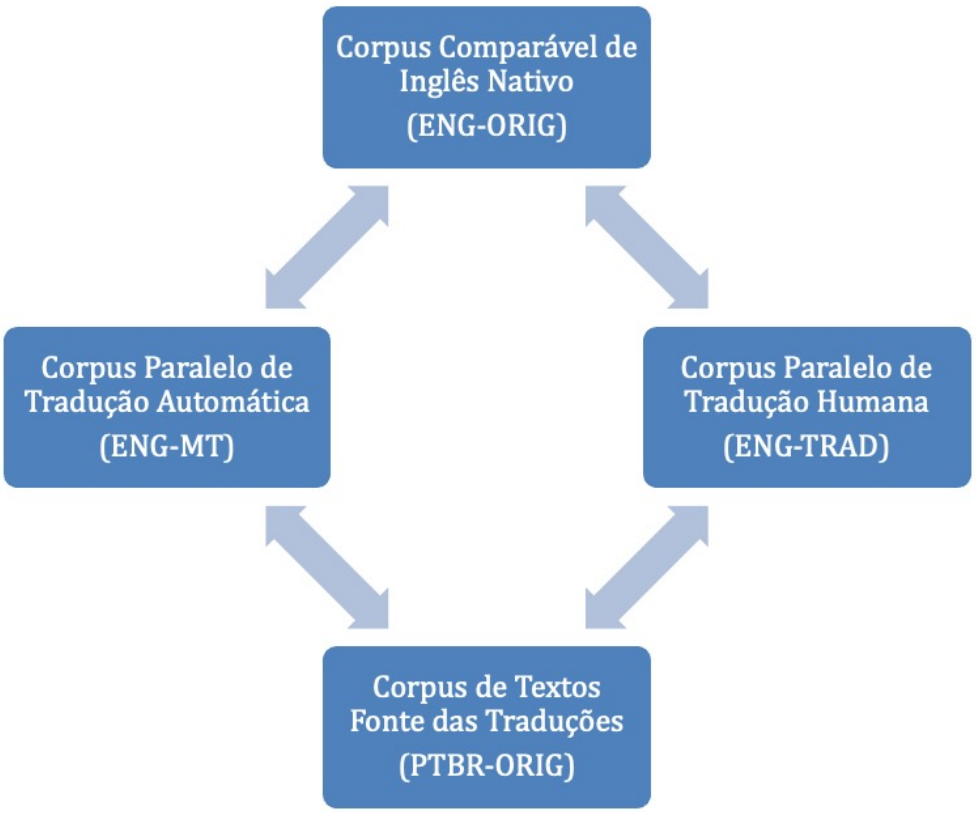
3 Metodologia

3.1 Design e coleta de corpus

O design e coleta dos corpora usados nesta pesquisa foram realizados de forma cuidadosa e minuciosa para garantir a representatividade dos textos selecionados e para possibilitar uma análise confiável dos dados obtidos.

Considerando os objetivos da pesquisa, definidos acima, foi definido que o corpus seria composto por três subcorpora diferentes, sendo dois corpora paralelos e um corpus comparável. Os corpora paralelos foram compostos de traduções do português para o inglês. O corpus comparável foi composto de textos provavelmente escritos por nativos da língua inglesa.

Figura 3 – Interrelação entre o corpus e os subcorpora



Fonte: Autoria própria.

Todos os textos de todos os subcorpora foram obtidos dos sites de empresas de capital aberto, divulgados e disponibilizados de forma online para livre acesso do público geral.

Abaixo, detalhamos, de forma geral, os três subcorpora usados nesta pesquisa e que fundamentaram as descobertas analisadas na discussão de resultados:

Tabela 1 - Visão geral do corpus da pesquisa

Registro	Subregistro		Língua	Nº Textos	Nº Palavras
	País de origem da companhia	Área de Atuação			
Comunicados e documentos corporativos de sociedades anônimas (empresas de capital aberto): comunicados ao mercado, fatos relevantes, avisos aos acionistas, estatutos sociais, políticas e atas de assembleias.	BRA	Educação	ENG-TRAD	100	37.956
			PTBR-ORIG		50.752
	EUA		ENG-MT		49.293
			ENG-ORIG	100	90.122
	BRA	Bancário	ENG-TRAD	100	77.953
			PTBR-ORIG		75.477
	EUA		ENG-MT		77.322
			ENG-ORIG	100	69.918
	BRA	Farmacêutica	ENG-TRAD	58	57.177
			PTBR-ORIG		57.033
	EUA		ENG-MT		57.412
			ENG-ORIG	100	209.280
	BRA	Software	ENG-TRAD	34	40.760
			PTBR-ORIG		41.255
	EUA		ENG-MT		40.667
			ENG-ORIG	100	89.704
	BRA	Pagamentos eletrônicos	ENG-TRAD	90	154.672
			PTBR-ORIG		155.376
	EUA		ENG-MT		154.944
			ENG-ORIG	54	35.491
	BRA	Energia	ENG-TRAD	100	91.958
			PTBR-ORIG		108.453
	EUA		ENG-MT		108.327
			ENG-ORIG	100	88.151
	BRA	Aviação	ENG-TRAD	100	99.996
			PTBR-ORIG		116.107
	EUA		ENG-MT		113.957
			ENG-ORIG	100	94.763
	BRA	Vestuário	ENG-TRAD	100	34.697
			PTBR-ORIG		34.979
	EUA		ENG-MT		34.368
			ENG-ORIG	100	76.796
	BRA	Resseguros	ENG-TRAD	100	59.227
			PTBR-ORIG		59.368
	EUA		ENG-MT		59.547
			ENG-ORIG	100	47.184
	BRA	Alimentos	ENG-TRAD	100	110.499
			PTBR-ORIG		116.944
	EUA		ENG-MT		118.160
			ENG-ORIG	100	66.033

Fonte: Autoria própria.

Tabela 2 – Número total de textos e de palavras do corpus da pesquisa

Língua	Total de Textos	Total de Palavras
ENG-TRAD	882	764.895
PTBR-ORIG	882	815.744
ENG-MT	882	813.997
ENG-ORIG	954	867.446
Total Geral	**3.600**	**3.262.082**

Fonte: Autoria própria.

- PTBR-ORIG = Textos provavelmente escritos por falantes nativos da língua portuguesa (português)
- ENG-TRAD = Tradução Humana-Oficial (português > inglês)
- ENG-MT = Tradução Automática (português > inglês)
- ENG-ORIG = Textos provavelmente escritos por falantes nativos da língua inglês (inglês)

3.1.1. Setores selecionados para o corpus

Tabela 3 – Visão geral dos setores de atuação do corpus da pesquisa

Setor	Características	Receita anual (Brasil)	Receita anual (Global)	Projeção de crescimento
Educação	- Fornece conhecimentos e habilidades - Promove aprendizagem e desenvolvimento - Promove igualdade de oportunidades	R$ 200 bilhões	US$ 6 trilhões	3%-6%
Bancário	- Fornece intermediação financeira - Presta serviços financeiros - Administra riscos financeiros	R$ 460 bilhões	US$ 4,7 trilhões	6%
Farmacêutica	- Pesquisa e desenvolve medicamentos e tratamentos - Atende a regulação e conformidade - Comercializa e distribui medicamentos	R$ 150 bilhões	US$ 1,4 trilhão	4%-6%
Softwares	- Foco em desenvolvimento e inovação tecnológica - Fornece personalização e adaptação - Fornece suporte e manutenção	R$ 131,6 bilhões	US$ 389 bilhões	8%
Pagamentos eletrônicos	- Processa pagamentos - Integra diversos meios de pagamentos - Administra riscos e segurança	R$ 400 bilhões	US$ 1,4 trilhão	12%
Energia	- Produz e distribui energia - Fornece diversas fontes de energia - Foco em sustentabilidade e eficiência energética	R$ 252 bilhões	US$ 5 trilhões	3%-4%
Aviação	- Fornece transporte aéreo - Foco em segurança e regulamentação aeroviária - Fornece conectividade global	R$ 53 bilhões	US$ 328 bilhões	0,3%

Vestuário	- Desenvolve e produz peças de vestuário - Distribui e comercializa tais peças - Foco em tendências de moda e estilo	R$ 210 bilhões	US$ 2,5 trilhões	3%
Resseguros	- Transfere riscos - Diversifica e avalia riscos - Foco em reaseguradores e retrocessão	R$ 23 bilhões	US$ 313 bilhões	3%
Alimentos	- Produz e processa alimentos - Foco em segurança alimentar e qualidade - Distribui e comercializa alimentos	R$ 800 bilhões	US$ 7,6 trilhões	4%

Fonte: Autoria própria.

3.1.2 Subregistros selecionados

Na tabela abaixo, os subregistros encontrados nos três corpora foram detalhados, juntamente com o quantitativo de amostras de acordo com cada subgênero.

Tabela 4 – Número de amostras do corpus separadas por subregistro

Tipo de subregistro	Número de amostras
ES	36
PO	269
CM	2.013
AA	501
FR	687
AT	114

Fonte: Autoria própria.

- ES – Estatuto Social: Documento legal que define a estrutura organizacional, direitos e responsabilidades dos acionistas, governança corporativa e regras de funcionamento dessas empresas.
- PO – Políticas: Diretrizes estabelecidas para governança, ética, responsabilidade social e práticas de negócios que orientam o comportamento e as decisões da empresa em sua atuação no mercado.
- CM – Comunicado/Aviso ao Mercado: Comunicados oficiais com informações relevantes e obrigatórias sobre a empresa, suas operações, eventos e decisões financeiras que podem impactar o mercado de ações e os investidores.
- AA – Comunicado/Aviso aos Acionistas/Detentores de Títulos de Crédito/Detentores de Debentures: Comunicados oficiais com informações importantes e obrigatórias sobre a empresa, suas atividades, resultados financeiros, decisões estratégicas e outros

assuntos relevantes, com o objetivo de manter a transparência e garantir os direitos dos acionistas/detentores de títulos de crédito/detentores de debentures.

- FR – Fato Relevante: Divulgações oficiais que informam ao mercado e aos acionistas sobre eventos, informações ou decisões significativas que podem impactar a empresa, ações e o mercado financeiro
- AT – Atas das Assembleias (Extraordinária/Ordinária): Registros oficiais que documentam as deliberações, decisões e resoluções tomadas durante as reuniões de acionistas, fornecendo um histórico formal das discussões e votações ocorridas.

3.1.3 Coleta e disponibilidade dos textos do corpus

Todos os textos dos corpora foram coletados dos sites de relações com investidores de cada uma das empresas acima. Um site de relações com investidores é uma plataforma digital criada e mantida por empresas de capital aberto para fornecer informações financeiras, operacionais e estratégicas relevantes aos investidores, acionistas e demais partes interessadas. O objetivo principal é promover a transparência, a comunicação eficiente e o acesso a informações que possam impactar as decisões de investimento.

Há diversos documentos disponíveis nesses sites, como balanços, demonstrações financeiras, relatórios de administração, comunicados e fatos relevantes, atas de assembleias e apresentações institucionais. Esses documentos são disponibilizados ao público para garantir a conformidade com as exigências dos órgãos reguladores, como a como a Comissão de Valores Mobiliários (CVM) no Brasil (COMISSÃO DE VALORES MOBILIÁRIOS, 2002) e a *Securities and Exchange Commission* (SEC) nos Estados Unidos (SECURITIES AND EXCHANGE COMMISSION, 2000), e para assegurar que todos os investidores e interessados tenham acesso equitativo às informações da empresa.

Além disso, ao disponibilizar esses documentos de forma aberta e transparente, a empresa demonstra seu compromisso com a governança corporativa e a ética empresarial, fortalecendo sua reputação e confiabilidade no mercado financeiro. Essa prática também permite que os investidores tomem decisões informadas e responsáveis, contribuindo para o sucesso e a sustentabilidade das empresas de capital aberto.

Nessa etapa, enfrentamos dois principais desafios:

I. Seleção das companhias brasileiras e americanas: não foi tarefa identificar companhias brasileiras e americanas equivalentes do mesmo setor, considerando que o ideal

seria que as empresas tivessem o mesmo tamanho, receitas anuais similares, níveis de governança idênticos, tivessem o mesmo escopo de atuação, entre outras características;

II. Disponibilidade de textos: foi necessário realizar uma verificação minuciosa de quais companhias atendiam os critérios do corpus em termos de materiais disponibilizados, textos equivalentes em português e inglês e número de textos disponibilizados. Dependendo do nível de governança ou do atual status da empresa (por exemplo, companhias em recuperação judicial), há diferentes níveis de obrigatoriedade em relação aos comunicados e documentos corporativos destas companhias. Empresas em recuperação judicial, por exemplo, não precisam traduzir todos os comunicados em português de forma imediata. Outra dificuldade foi o tempo de operação da companhia. Empresas mais jovens, listadas há menos tempo na bolsa de valores, têm um número infinitamente menor de documentos disponibilizados. Esse desafio implicou em substituir diversas companhias selecionadas no começo da pesquisa. No setor bancário, por exemplo, foi necessário trocar a companhia selecionada três vezes, no meio da coleta do corpus, pois algum dos empecilhos acima impediram a continuidade de coleta da companhia inicialmente escolhida.

3.2 Corpora paralelos e corpus comparável

Os Estudos da Tradução baseados em corpus empregam corpora paralelos e comparáveis, permitindo aos pesquisadores explorar as relações e diferenças entre várias línguas. Esses recursos são amplamente utilizados para analisar e aperfeiçoar processos de tradução, desenvolver ferramentas de tradução automática e investigar fenômenos linguísticos. Como explicado por Kenny "Embora esse trabalho esteja apenas começando agora, espera-se que a constante avaliação das categorias analíticas nos estudos de tradução com base nos dados contidos em corpora paralelos leve a uma maior clarificação conceitual no campo dos estudos de tradução como um todo[8]". (KENNY, 2006, p. 51)

Os corpora paralelos são coleções de textos em que cada documento de origem é acompanhado por sua tradução em outra língua. Esses recursos são particularmente úteis para estudar padrões de tradução e identificar equivalências linguísticas (BAKER, 1995). Além disso, os corpora paralelos têm sido fundamentais para o desenvolvimento de sistemas de

[8] Original: *Although such work is only now beginning, it is hoped that the constant testing of analytical categories in translation studies against data contained in Parallel corpora, will lead to greater conceptual clarification in translation studies as a whole*.

tradução automática, que utilizam técnicas de aprendizagem por máquina para extrair padrões a partir desses exemplos alinhados.

Os corpora comparáveis, por outro lado, são conjuntos de textos em duas ou mais línguas que compartilham características temáticas, de gênero ou estilo, mas não são necessariamente traduções uns dos outros. Esses corpora possibilitam análises contrastivas entre as línguas, permitindo que os pesquisadores identifiquem semelhanças e diferenças estruturais, lexicais e gramaticais. Resende (2019, p. 38) esclarece que "O interessante sobre os corpora comparáveis é que eles, como o nome diz, podem ser comparados, porque são organizados sob os mesmos princípios [...] geralmente com o mesmo tópico e assunto. Dessa forma, os corpora comparáveis contribuem para uma compreensão mais aprofundada dos fenômenos interlinguísticos e das estratégias de tradução.

3.2.1 Corpora paralelos

O uso de corpora paralelos na pesquisa de Linguística de Corpus e tradução envolve o estudo de dois ou mais corpora que contenham textos equivalentes em diferentes idiomas, como o corpus de tradução automática e o corpus de tradução humana. Esses corpora paralelos podem ser usados para comparar as traduções geradas por diferentes métodos e analisar a qualidade e as características linguísticas das traduções em diferentes idiomas (BAKER, 1995).

A comparação entre os corpora paralelos permite aos pesquisadores identificar padrões e tendências que podem ser atribuídos a características específicas dos métodos de tradução, como o uso de algoritmos para a tradução automática e a intervenção humana na tradução humana (LAVIOSA, 2002). Nesta pesquisa, comparamos um corpus de tradução automática com um corpus de tradução humana.

Uma abordagem comum para analisar e comparar corpora paralelos é utilizar medidas estatísticas e técnicas de análise de texto, como a análise de correspondência de termos, a análise de frequência de palavras e estruturas gramaticais (BAKER, 1995). Essas análises podem ajudar a identificar diferenças e semelhanças entre as traduções geradas pelos diferentes métodos e fornecer informações valiosas sobre como melhorar a qualidade das traduções.

Ao comparar o corpus de tradução automática com o corpus de tradução humana, podemos identificar áreas onde a tradução automática pode ser aprimorada e onde os tradutores humanos podem se beneficiar do uso de ferramentas de tradução assistida por computador (WAY, 2018). Além disso, as análises de corpora paralelos podem ser usadas para melhorar a

compreensão dos processos cognitivos envolvidos na tradução e para contribuir com o desenvolvimento de melhores práticas de tradução (BAKER, 1995).

A utilização de corpora paralelos na pesquisa de Linguística de Corpus e tradução é uma abordagem poderosa para explorar e entender as diferenças e semelhanças entre as traduções geradas por diferentes métodos. Essas análises fornecem informações valiosas para o aprimoramento contínuo das tecnologias de tradução e para o desenvolvimento de estratégias de tradução mais eficazes e precisas.

3.2.1.1 Corpus de tradução humana-oficial

O corpus de tradução humana-oficial é composto de traduções oficiais obtidas nos sites das companhias de capital aberto e disponibilizadas ao público. Companhias listadas em bolsas de valores internacionais são obrigadas a disponibilizar todo conteúdo de relações com investidores em português (original) e em inglês (versão traduzida).

Embora a qualidade das traduções disponibilizadas nos sites de empresas de capital aberto seja de extrema importância para garantir a compreensão correta das informações apresentadas, não há garantia de que todas essas traduções sejam feitas por tradutores humanos. Um dos motivos para isso é o desenvolvimento e a popularização de sistemas de tradução automática, como o Google Tradutor (WU et al., 2016) e o DeepL (ROLLER et al., 2020), que têm demonstrado um desempenho cada vez mais próximo ao nível humano em várias tarefas de tradução, uma das motivações para os objetivos desta pesquisa. Tais sistemas são frequentemente utilizados por empresas para economizar tempo e custos, especialmente quando o volume de informações a ser traduzido é extenso.

Outro fator que pode levar à utilização de tradução automática em vez de tradutores humanos é a crescente demanda por traduções em tempo real, em um cenário onde a informação é constantemente atualizada e os mercados globais exigem respostas rápidas (O'BRIEN, 2012). De fato, a tradução automática tem sido usada em contextos corporativos em que a velocidade é mais importante do que a qualidade da tradução (GASPARI, ALMAGHOUT & DOHERTY, 2014). No entanto, é importante salientar que, mesmo quando a tradução automática é utilizada, a presença de um revisor humano pode ser fundamental para garantir a precisão e a adequação das informações traduzidas, especialmente em contextos financeiros e corporativos onde erros de comunicação podem ter consequências significativas (KENNY & DOHERTY, 2014).

Além disso, mesmo que as traduções disponibilizadas nos sites de empresas de capital aberto tenham sido realizadas por tradutores humanos, não há garantia de que esses

profissionais sejam especializados na área de tradução financeira. A tradução financeira é um campo altamente especializado, que exige conhecimento profundo de terminologias específicas, bem como familiaridade com as práticas e regulamentações do setor (GARCÍA, 2015). No entanto, devido à demanda crescente por serviços de tradução, muitas vezes as empresas podem contratar tradutores sem a devida especialização, seja por falta de disponibilidade de especialistas ou por questões orçamentárias (PYM et al., 2012).

A falta de padronização na formação e na certificação de tradutores profissionais também contribui para a incerteza quanto à qualificação dos profissionais envolvidos (DRUGAN, 2013). Embora existam organizações e associações profissionais, como a *American Translators Association* (ATA) e a Associação Brasileira de Tradutores e Intérpretes (ABRATES), que oferecem certificações específicas para tradutores, nem todos os tradutores que trabalham no setor possuem tais certificações (ABDALLAH, 2012). Dessa forma, não há garantias de que as traduções disponíveis nos sites de empresas de capital aberto tenham sido realizadas por tradutores qualificados e especializados na área financeira, o que pode afetar a precisão e a confiabilidade das informações traduzidas.

Apesar das incertezas quanto à origem das traduções disponibilizadas nos sites de empresas de capital aberto, é razoável presumir que muitas delas sejam feitas por tradutores humanos devidamente qualificados e especializados, devido à importância e às implicações legais das informações divulgadas por essas empresas. A legislação e as normas de governança corporativa exigem que as empresas de capital aberto forneçam informações claras, precisas e compreensíveis aos seus investidores e partes interessadas (*Securities and Exchange Commission,* 2000). Isso coloca uma responsabilidade significativa sobre as empresas em relação à qualidade das traduções disponíveis em seus sites, levando-as a priorizar a contratação de tradutores humanos qualificados para garantir a precisão e a adequação das informações traduzidas (GARCÍA, 2015).

Além disso, é importante considerar que, embora os sistemas de tradução automática tenham avançado significativamente nos últimos anos, eles ainda enfrentam desafios ao lidar com textos altamente especializados, como os encontrados na área financeira. Tradutores humanos, por outro lado, têm a capacidade de compreender e processar as nuances, ambiguidades e terminologias específicas presentes nesses textos, o que é essencial para garantir a qualidade das traduções (GARCÍA, 2015). Assim, é provável que muitas empresas de capital aberto optem por contratar tradutores humanos para lidar com as suas traduções, a fim de cumprir as exigências regulatórias e garantir a qualidade das informações disponibilizadas aos investidores e partes interessadas.

As seguintes etapas foram realizadas na coleta do corpus de tradução humana:

a) Os textos selecionados foram baixados no formato PDF diretamente dos sites de Relações com Investidores em inglês de cada uma das empresas selecionadas, garantindo que cada um dos textos tinha um texto correspondente em português;
b) Os textos foram convertidos do PDF para o Word utilizando a ferramenta Adobe Acrobat;
c) Os textos convertidos passaram por um cotejo com os documentos originais em PDF para verificar e corrigir erros de conversão;
d) Todos os textos passaram por uma etapa de limpeza da formatação, na qual removemos palavras/frases em negrito, sublinhadas e/ou em itálico e o texto foi padronizado com uma única fonte no mesmo tamanho;
e) Os textos foram renomeados para facilitar a localização dos textos no corpus;
f) Por fim, os textos traduzidos foram convertidos do formato Word para o formato TXT.

3.2.1.2 Corpus de tradução automática

Neste estudo, além da coleta de traduções oficiais nos sites de empresas de capital aberto, focamos na produção e coleta de um corpus de tradução automática, com o objetivo de analisar e avaliar o desempenho dos sistemas de tradução automática em diferentes contextos e identificar áreas que possam ser aprimoradas. A crescente popularidade e o desenvolvimento contínuo de sistemas de tradução automática, como o Google Tradutor e o DeepL, tornam esse tema de pesquisa particularmente relevante e oportuno. Ao estudar o corpus de tradução automática, buscamos compreender os padrões, tendências e desafios enfrentados por essas ferramentas na tradução de textos de diversos gêneros e domínios, o que pode contribuir para o aperfeiçoamento desses sistemas.

As seguintes etapas foram realizadas na coleta do corpus de tradução automática:

a) Os textos selecionados foram baixados no formato PDF diretamente dos sites de Relações com Investidores em português de cada uma das empresas selecionadas, garantindo que cada um dos textos tinha um texto correspondente em inglês;
b) Os textos foram convertidos do PDF para o Word utilizando a ferramenta Adobe Acrobat;

c) Os textos convertidos passaram por um cotejo com os documentos originais em PDF para verificar e corrigir erros de conversão;
d) Todos os textos passaram por uma etapa de limpeza da formatação, na qual removemos palavras/frases em negrito, sublinhadas e/ou em itálico e o texto foi padronizado com uma única fonte no mesmo tamanho;
e) Os textos foram renomeados para facilitar a localização dos textos no corpus;
f) Os textos foram inseridos na ferramenta memoQ;
g) A API de tradução automática instalada na ferramenta foi executada em todos os arquivos para traduzir os textos do português para o inglês;
h) Os segmentos traduzidos foram verificados para remover possíveis tags remanescentes mesmo após o processo de limpeza da formatação;
i) Os segmentos traduzidos foram confirmados na memória de tradução e os arquivos foram exportados da ferramenta;
j) Por fim, os textos traduzidos foram convertidos do formato *Word* para o formato TXT.

Abaixo, abordaremos com mais detalhes o uso do memoQ, uma das ferramentas de tradução assistida por computador (CAT) mais populares entre os profissionais da área e ferramenta escolhida para os objetivos desta pesquisa. Também apresentaremos o recurso de tradução automática integrado a essa plataforma e a utilização desse recurso para gerar as versões traduzidas do corpus analisado.

I. CAT Tool (MemoQ)

O memoQ é uma ferramenta de tradução assistida por computador amplamente utilizada, projetada para melhorar a eficiência e a qualidade do trabalho dos tradutores. Desenvolvida pela Kilgray Translation Technologies, oferece recursos avançados de gerenciamento de projetos, memória de tradução, terminologia e integração com sistemas de tradução automática. A memória de tradução (TM) é um recurso essencial que armazena segmentos de texto previamente traduzidos para reutilização futura, garantindo consistência e economizando tempo. O gerenciamento de terminologia possibilita a criação e manutenção de glossários específicos do projeto, mantendo a precisão terminológica. A integração com sistemas de tradução automática permite aos tradutores usar essas tecnologias como ponto de partida para suas traduções, combinando a capacidade da tradução automática com a perícia humana.

O memoQ também oferece recursos avançados de gerenciamento de projetos, permitindo atribuir tarefas a vários tradutores, revisores e especialistas em terminologia, facilitando a colaboração e a comunicação entre a equipe. As tags desempenham um papel crucial no processo de tradução assistida por computador, preservando a formatação e a estrutura dos documentos originais. Elas são inseridas no texto-fonte e indicam elementos não textuais, como formatação, hiperlinks e quebras de linha. O memoQ facilita o gerenciamento de tags, permitindo que sejam inseridas automaticamente no texto de destino e oferecendo recursos de validação para evitar problemas de formatação.

No entanto, o uso de tags pode levar a poluição visual no texto, dificultando a leitura e a compreensão do conteúdo pelos tradutores. Além disso, é necessário um tempo de aprendizado para o uso adequado das tags e os recursos disponíveis para gerenciá-las eficientemente. Na presente pesquisa, as tags representaram uma dificuldade adicional na coleta adequada do corpus de tradução automática, exigindo etapas adicionais para eliminar as tags geradas.

A integração da API da Tradução Neural Avançada do Google Tradutor no memoQ permite aos tradutores utilizar recursos de tradução automática avançada diretamente na interface. Essa tradução neural avançada é baseada em redes neurais artificiais e algoritmos de aprendizado profundo, proporcionando traduções mais naturais e precisas. Os tradutores podem enviar segmentos de texto selecionados ao Google Tradutor por meio da API, recebendo traduções automáticas em tempo real que podem ser revisadas e aprimoradas conforme necessário.

Em resumo, o memoQ é uma ferramenta completa para tradução assistida por computador, oferecendo recursos avançados para melhorar a eficiência e a qualidade das traduções. Sua memória de tradução, gerenciamento de terminologia, integração com tradução automática e recursos de gerenciamento de projetos facilitam o trabalho dos tradutores e promovem a colaboração na equipe. As tags desempenham um papel importante na preservação da formatação e da estrutura dos documentos, embora exijam cuidados na sua gestão. A integração da API da Tradução Neural Avançada do Google Tradutor proporciona.

II. Tradução Automática pelo Google Tradutor

O Google Tradutor emprega tecnologia que utiliza redes neurais artificiais e algoritmos de aprendizado profundo para oferecer traduções mais precisas e naturais em comparação com os métodos tradicionais. Ela é treinada em grandes volumes de texto paralelo e consegue lidar

com várias construções gramaticais, inclusive expressões idiomáticas e coloquiais. No entanto, a ferramenta pode apresentar dificuldades com ambiguidades e a qualidade das traduções pode variar entre idiomas.

Embora a tradução automática neural avançada tenha suas limitações, representa um avanço significativo na área de tradução automática. Espera-se que continue a melhorar à medida que mais dados de treinamento e avanços algorítmicos se tornem disponíveis. A colaboração entre desenvolvedores, linguistas e tradutores profissionais também desempenha um papel importante na melhoria contínua da ferramenta, garantindo que ela possa lidar com nuances culturais e linguísticas específicas.

É fundamental ressaltar que, apesar dos avanços na tradução automática neural avançada, ela não substitui completamente os tradutores humanos, especialmente em situações que exigem precisão, sensibilidade cultural ou conhecimento especializado. A ferramenta deve ser vista como uma ferramenta complementar aos serviços de tradução profissional, ajudando os tradutores a trabalhar de maneira mais eficiente e aprimorar a qualidade de suas traduções.

Em resumo, a tradução automática neural avançada do Google Tradutor tem o potencial de aprimorar significativamente a comunicação global e a compreensão entre culturas e idiomas. Embora haja desafios a serem superados, a contínua evolução da tecnologia e a colaboração entre as partes interessadas garantem um futuro promissor para a tradução automática em uma ampla gama de contextos.

3.2.2 Corpus comparável

O uso de corpora comparáveis na pesquisa de Linguística de Corpus e tradução envolve a análise de corpora distintos, cada um representando diferentes tipos de textos, a fim de investigar padrões linguísticos, variações e características específicas de cada corpus. Neste caso específico, o corpus comparável foi comparado ao corpus de tradução automática e ao corpus de tradução humana para explorar como os processos de tradução automática e humana afetam a qualidade e as características linguísticas das traduções (BAKER, 1993).

A comparação entre esses três corpora permite aos pesquisadores identificar padrões e tendências que podem ser atribuídos às características específicas do método de tradução. Por exemplo, a tradução automática pode apresentar inconsistências ou erros sistemáticos devido à sua abordagem algorítmica, enquanto a tradução humana pode ser mais propensa a variações individuais devido às escolhas e preferências dos tradutores (LAVIOSA, 1998).

Uma abordagem comum para analisar e comparar esses corpora é utilizar medidas estatísticas e técnicas de análise de texto, como a análise de frequência de palavras, colocações, n-gramas e estruturas gramaticais (OLOHAN, 2004). Essas análises podem revelar diferenças e semelhanças nos padrões de uso da língua entre os corpora e ajudar os pesquisadores a entender como a tradução automática e humana impactam o resultado da tradução.

Ao comparar o corpus de textos provavelmente escritos por falantes nativos do inglês com o corpus de tradução automática e o corpus de tradução humana, o objetivo foi identificar áreas onde a tradução automática pode ser aprimorada e onde os tradutores humanos podem se beneficiar do uso de ferramentas de tradução assistida por computador (WAY, 2018). Além disso, essas análises podem fornecer informações valiosas para o desenvolvimento de melhores algoritmos de tradução automática e para o treinamento e aperfeiçoamento de tradutores profissionais (GASPARI, ALMAGHOUT & DOHERTY, 2014).

Em conclusão, a utilização de corpora comparáveis na pesquisa de Linguística de Corpus e tradução é uma abordagem poderosa para explorar e entender as diferenças e semelhanças entre traduções automáticas e humanas em relação ao texto original. Essas análises fornecem informações valiosas para o aprimoramento contínuo das tecnologias de tradução e para o desenvolvimento de estratégias de tradução mais eficazes e precisas.

As seguintes etapas foram realizadas na coleta do corpus de textos provavelmente escritos por falantes nativos do inglês:

a) Os textos selecionados foram baixados no formato PDF diretamente dos sites de Relações com Investidores em inglês de cada uma das empresas selecionadas;
b) Os textos foram convertidos do PDF para o Word utilizando a ferramenta Adobe Acrobat*;*
c) Os textos convertidos passaram por um cotejo com os documentos originais em PDF para verificar e corrigir erros de conversão;
d) Todos os textos passaram por uma etapa de limpeza da formatação, na qual removemos palavras/frases em negrito, sublinhadas e/ou em itálico e o texto foi padronizado com uma única fonte no mesmo tamanho;
e) Os textos foram renomeados para facilitar a localização dos textos no corpus;
f) Por fim, os textos foram convertidos do formato Word para o formato TXT.

3.2.3. Dificuldades na conversão dos textos do corpus

No início da pesquisa, não prevemos dificuldades na conversão de documentos obtidos com as companhias de capital aberto. Os softwares de conversão de imagem em texto, também conhecidos como *Optical Character Recognition* (OCR) têm se mostrado cada vez mais sofisticados e eficientes. Esses avanços tecnológicos têm facilitado consideravelmente o trabalho de profissionais de diversas áreas ao permitir uma conversão mais precisa e eficiente, contribuindo para uma análise mais precisa e confiável dos dados linguísticos.

No entanto, essa conversão foi o principal desafio enfrentado na coleta do corpus. A conversão era uma etapa considerada simples no projeto, com previsão de levar somente alguns dias para ficar pronta. Contudo, foram necessários quase três meses para converter os documentos de forma satisfatória para que não comprometesse a análise de dados. Para isso, após a conversão, todos os textos passaram por uma limpeza e cotejo. Além disso, esses erros de conversão levaram um tempo para ser notados, tendo em vista que passavam desapercebidos e só se tornaram óbvios no processo de inserção na CAT Tool ou da conversão em TXT.

Abaixo, listamos os principais erros de conversão que necessitaram de correção:

I. Tags geradas no MemoQ
II. Segmentação incorreta do texto
III. Omissão de trechos do texto
IV. Diferentes fontes na mesma frase ou trecho
V. Tags geradas no documento TXT
VI. Falha na conversão de certas palavras, majoritariamente palavras com acentuação e/ou Ç

Os erros listados, além de gerar problemas na leitura e processamento de dados no SAS e no Weka, também comprometem a fidedignidade e eficiência da ferramenta de tradução automática. Isso ocorre principalmente em trechos com falhas na conversão de palavras e textos com segmentações incorretas.

3.3. Processamento do corpus

3.3.1 Etiquetagem

A etiquetagem do corpus é um processo fundamental na análise multidimensional da Linguística de Corpus. Esse processo envolve a marcação de diferentes elementos linguísticos nos textos, como palavras, frases, expressões e estruturas gramaticais, para que possam ser facilmente identificados e analisados (BERBER SARDINHA, 2004).

Existem diferentes técnicas e ferramentas que podem ser utilizadas para realizar a etiquetagem morfossintática do corpus – em inglês, *Part-of-Speech* (POS) *tagging* –, incluindo o uso de softwares especializados que identificam a classe gramatical de cada palavra em um texto (MCENERY & HARDIE, 2011). Outras técnicas incluem a etiquetagem manual, realizada por linguistas especializados, e a etiquetagem semiautomática, que combina a intervenção humana com o uso de algoritmos computacionais. Neste estudo em específico, utilizamos o Biber Tagger (BIBER, 1988, 1995; BERBER SARDINHA & VEIRANO PINTO, 2014, 2019) para etiquetar os corpora.

3.3.2 SAS OnDemand

O SAS *OnDemand for Academics* é uma plataforma baseada em nuvem que permite o acesso a diferentes ferramentas e recursos para gestão, análise e apresentação de dados, incluindo análise de texto em larga escala (SAS, 2021). O SAS *OnDemand for Academics* é uma versão do software disponível gratuitamente na rede, que dispensa a necessidade de instalação e inclui vários módulos de análise estatística.. Os usuários podem importar diferentes tipos de dados e formatos de arquivo, como planilhas, arquivos de texto e bancos de dados (SAS, 2021).

A seguir, introduzimos o etiquetador morfossintático Biber Tagger.

I. Biber Tagger

O Biber Tagger, desenvolvido pelo linguista Douglas Biber (BIBER, 1988, 1995), é uma ferramenta de etiquetagem de palavras utilizada na pesquisa de Linguística de Corpus. Ele é capaz de identificar e marcar automaticamente diversas características linguísticas da língua inglesa, tais como categorias de natureza morfossintática.

Por exemplo, em um estudo que investiga a tradução de gêneros ou registros do inglês para o português, o Biber Tagger pode ser usado para identificar as dimensões presentes nos gêneros ou registros do corpus em inglês e comparar sua frequência de uso entre as diferentes dimensões de variação linguística em outras línguas – por exemplo, Berber Sardinha, Kauffmann e Acunzo (2014), cuja pesquisa conduzida analisou diversos gêneros ou registros em português sob o prisma das dimensões, utilizando a mesma metodologia da análise multidimensional de Biber (1988). Isso permite aos pesquisadores observarem como os

tradutores lidam com esses gêneros ou registros, e se há alguma variação na tradução de acordo com o contexto em que se posicionam as dimensões de variação nas línguas-alvo.

Além disso, o Biber Tagger também é útil na identificação de padrões dimensionais de coocorrência em diversos registros ou gêneros, o que pode ajudar a entender como as palavras são usadas em diferentes contextos. Por exemplo, em um estudo sobre a linguagem de negociação em inglês, o *Biber Tagger* pode ser usado para identificar quais dimensões de variação linguística influenciam palavras e estruturas sintáticas mais frequentes em negociações bem-sucedidas.

De acordo com Biber (1995), o uso de corpora é fundamental para a pesquisa em variação linguística, uma vez que permite aos pesquisadores observarem como a linguagem é usada em diferentes contextos e, assim, fazerem generalizações sobre sua natureza e uso. O Biber Tagger é uma ferramenta importante nesse processo, permitindo que os pesquisadores processem e analisem um grande volume de palavras da língua inglesa de forma eficiente e precisa.

3.3.3 Análise lexical com o Weka

O Weka (*Waikato Environment for Knowledge Analysis*) é uma plataforma popular de mineração de dados e aprendizagem por máquina desenvolvida pela Universidade de Waikato, na Nova Zelândia (HOLMES, DONKIN & WITTEN, 1994). Com sua interface gráfica de fácil uso, o Weka permite a análise lexical por meio da extração de características linguísticas e aplicação de algoritmos de aprendizagem por máquina, como J48 e Random Forest, para classificação, agrupamento e previsão de resultados.

No contexto da Linguística de Corpus, o Weka pode ser utilizado para análise de padrões de frequência de palavras e construções sintáticas, possibilitando identificar características mais frequentes em determinados contextos, como a linguagem de negociação em inglês. Além disso, o Weka tem sido explorado em estudos que buscam criar modelos de classificação capazes de diferenciar traduções automáticas e humanas, por meio do treinamento de algoritmos de aprendizagem por máquina.

A aprendizagem por máquina é uma subárea da inteligência artificial que visa desenvolver algoritmos capazes de aprender a partir dos dados, sem programação explícita. Nesse sentido, o Weka se destaca ao permitir a coleta, organização e análise de dados, identificando padrões e criando modelos de classificação baseados em técnicas estatísticas e matemáticas.

No Weka, é possível utilizar conjuntos de dados de treinamento contendo traduções automáticas e humanas para criar modelos de classificação que ajudam a diferenciar as categorias (HOLMES, DONKIN & WITTEN, 1994). Com base em algoritmos de aprendizagem supervisionada, o modelo é treinado com exemplos rotulados e, posteriormente, é capaz de classificar novas traduções, contribuindo para a avaliação da qualidade das traduções automáticas e humanas, além da detecção de possíveis problemas.

Para processar os dados com o Weka, realizamos as seguintes etapas:

Em primeiro lugar, selecionamos o conjunto de dados utilizados para treinar o modelo de classificação. Esse conjunto de dados contém exemplos rotulados, ou seja, cada exemplo tem uma categoria associada.

Em seguida, o conjunto de dados foi dividido em conjuntos de treinamento e teste. O conjunto de treinamento foi utilizado para treinar o modelo de classificação, enquanto o conjunto de teste foi utilizado para avaliar a precisão do modelo.

Após treinar o modelo de classificação, utilizamos o algoritmo para analisar novos dados e classificá-los de acordo com as categorias definidas.

4 Resultados

Para prosseguirmos aos resultados da pesquisa, é necessário relembrar os dois principais objetivos desta dissertação e análise:

1) Descobrir se há traços linguísticos que podem diferenciar de forma probabilística a tradução humana da tradução automática; e

2) Identificar possíveis traços linguísticos que podem diferenciar de forma probabilística e estatística uma tradução para o inglês (automática ou humana) de um texto escrito em inglês.

Sendo assim, o processamento do corpus, conforme a metodologia detalhada previamente, gerou os resultados abaixo descritos.

4.1 Análise lexical com Weka

4.1.1 Algoritmo Random Forest

Por meio do processamento com o algoritmo Random Forest, a análise lexical via Weka comparou o corpus de tradução automática e o corpus de tradução humana. Tal comparação gerou 529 amostras, das quais 445 amostras foram classificadas corretamente e 84 amostras foram classificadas incorretamente.

O corpus total de 1.762 amostras foi dividido da seguinte forma: 70% (1.233 amostras) foram utilizadas para construir a base de conhecimento do algoritmo por meio da aprendizagem por máquina, o que gerou 481 atributos para análise. Após desenvolver essa base, o Random Forest testou as 529 amostras restantes.

O índice de acerto foi de 84,12%. Tal índice significa que o algoritmo, ao processar e analisar uma tradução, pode identificar corretamente em 8 a cada 10 traduções se essa tradução foi feita de forma automática (ou seja, traduzida por máquina) ou se essa tradução foi feita com um profissional da área de tradução.

Tabela 5 – Visão geral dos resultados do algoritmo Random Forest

Amostras Classificadas Corretamente	445	82,121%
Amostras Classificadas Incorretamente	84	15,879%
Número Total de Amostras	529	100,000%

Fonte: Autoria própria.

Na tabela abaixo, temos a matriz de confusão da classificação.

Tabela 6 – Matriz de confusão

	Classificação inicial	
Classificação por máquina	Tradução Humana	Tradução automática
Tradução Humana	229	44
Tradução automática	40	216
Total	269	260

Fonte: Autoria própria.

4.1.2 Algoritmo J48

Fizemos duas análises diferentes com o algoritmo J48.

Na primeira análise, utilizamos as mesmas porcentagens do Random Forest, ou seja, 70% das amostras foram utilizadas para treino e 30% efetivamente testadas. Sendo assim, do total de 1.762 amostras, tivemos 529 amostras testadas pelo algoritmo.

Considerando essas porcentagens, tivemos os seguintes resultados: 439 amostras foram classificadas corretamente e somente 90 amostras foram classificadas incorretamente. O índice de acerto foi de 82,99% e o índice de erro foi de 17,01%.

Tabela 7 - Visão geral dos resultados do algoritmo J48 (70% treino e 30% teste)

Amostras Classificadas Corretamente	439	82,99%
Amostras Classificadas Incorretamente	90	17,01%
Número Total de Amostras	529	100%

Fonte: Autoria própria.

Na segunda análise com o algoritmo J48, dividimos as 1.762 amostras de forma diferente quando comparado ao algoritmo *Random Forest*. Do total de amostras, 1.175 foram utilizadas na aprendizagem por máquina para reconhecer os atributos de cada uma das traduções parte dos corpora, perfazendo 60% das amostras. Os 40% restantes foram testados com base nos atributos reconhecidos previamente.

O processamento com o algoritmo J48 gerou resultados similares à análise com o algoritmo *Random Forest*. A comparação entre o corpus de tradução automática e o corpus de

tradução humana gerou 705 amostras com 587 classificados corretamente e 118 classificadas incorretamente. O índice de previsibilidade probabilística com este método foi de 83,26%, ligeiramente inferior ao método Random Forest, mas ainda assim um índice significativo, principalmente ao considerarmos o volume de amostras.

Tabela 8 - Visão geral dos resultados do algoritmo J48 (60% treino e 40% teste)

Amostras Classificadas Corretamente	587	83,26%
Amostras Classificadas Incorretamente	118	16,74%
Número Total de Amostras	705	100%

Fonte: Autoria própria.

As tabelas abaixo detalham os índices de acerto do corpus de tradução automática e do corpus de tradução humana para o algoritmo J48, tanto para a análise com 70% de treino e 30% de teste, quanto para a análise com 60% de treino e 40% de teste.

Tabela 9 – Matriz de confusão

	Classificação inicial	
Classificação por máquina	Tradução humana	Tradução automática
Tradução humana	240	33
Tradução automática	57	199
Total	297	232

Fonte: Autoria própria.

Tabela 10 - Composição das amostras e índices de acerto dos resultados do algoritmo J48 (60% treino e 40% teste)

	Classificação inicial	
Classificação por máquina	Tradução humana	Tradução automática
Tradução humana	292	63
Tradução automática	55	295
Total	347	358

Fonte: Autoria própria.

Na primeira tabela, com J48 - 70% treino, 30% teste - é possível observar que o algoritmo foi mais eficaz (+10,18%) em identificar as traduções automáticas quando comparado ao índice de acerto das traduções humanas. Já na segunda tabela com J48, sendo 60% de treino e 40% de teste, o número de amostras classificadas erroneamente é similar entre os corpora.

Conforme os exemplos abaixo que evidenciam tal conclusão, com base nos dados, essa discrepância pode indicar que, nesta análise em específico, tivemos muitos marcadores lexicais de tradução automática presentes na tradução humana-oficial. Outra possibilidade neste cenário é que os indicadores lexicais para tradução automática foram ligeiramente menos assertivos, apesar do alto índice de acerto de forma geral.

No entanto, é possível observar que as análises puramente via Weka, ou seja, Random Forest (70% treino, 30% teste) e J48 (60% treino, 40% teste) tiveram resultados padronizados e consistentes em relação aos erros e acertos, indicando que os marcadores lexicais foram mais precisos e assertivos que a análise com o algoritmo J48 com 70% de treino e 30% de teste.

Na análise Weka com o algoritmo J48 (70% treino, 30% teste), os seguintes indicadores, apresentados na Tabela 11, sobressaem como marcadores linguísticos para realizar a identificação e diferenciação mencionadas no parágrafo anterior.

Tabela 11 - Marcadores linguísticos

Tradução Automática		**Tradução Humana-Oficial**	
Palavra	Classificação Gramatical	Palavra	Classificação Gramatical
company	substantivo	*held*	verbo
a	preposição	*information*	substantivo
resolution	substantivo	*officer*	substantivo
vice	adjetivo	*executive*	adjetivo

Fonte: Autoria própria.

Com base na comparação dos corpora com os métodos de análise lexical, é possível concluir, ao combinar os resultados do Weka com a leitura e conhecimento dos textos que compõe os corpora estudados, que há alguns itens lexicais fundamentais para a identificação correta das traduções e o alto índice de acerto, conforme os exemplos abaixo.

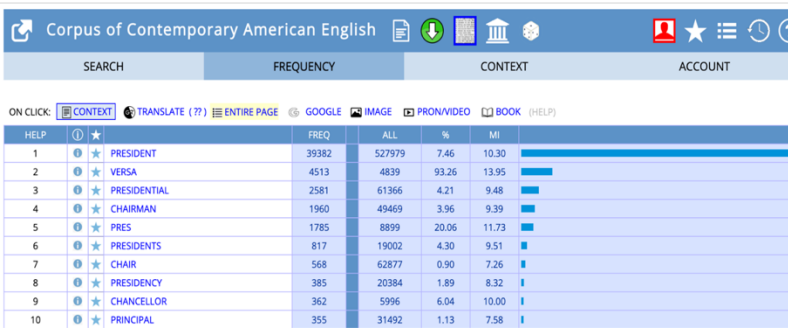
I. Colocações

Conforme colocado, colocações são combinações de palavras que ocorrem de forma frequente em um idioma, com um significado específico. Podemos destacar dois exemplos presentes na Tabela 11 de marcadores linguísticos lexicais:

No corpus de tradução automática, um dos indexadores foi o adjetivo "*vice*". Na tradução automática, a palavra "vice-presidente" é comumente traduzida como "*vice president*". Importante ressaltar que, gramaticalmente e contextualmente, essa tradução não está incorreta. A hifenização é opcional.

Na imagem abaixo, podemos observar que há uma enorme frequência da colocação "*president*" associada ao termo "*vice*", de acordo com o Corpus do Inglês Americano Contemporâneo (COCA - *Corpus of Contemporary American English*).

Figura 4 – Tipo e frequência de colocações para o termo "*vice*" no COCA Corpus



Fonte: https://www.english-corpora.org/coca/

No entanto, em traduções e textos da área financeira produzidos por falantes cuja língua materna é o inglês, principalmente no contexto do inglês americano, a prolixidade é sempre evitada e a abreviação é extremamente comum, quando não de praxe. Portanto, esta palavra é muito frequente traduzida simplesmente com "VP" por tradutores experientes e especializados.

Ao comparar a frequência da colocação "*vice president*" com a frequência do uso da palavra VP, nota-se uma frequência muito maior do termo "VP", conforme a tabela abaixo, com base no Corpus de Inglês Americano Contemporâneo.

Tabela 12 – Comparação de termos com base em frequência no COCA Corpus

	Vice-President	**Vice President**	**VP**
Frequência	1938	1993	4793
Palavras (M)	993	993	993

Por Milhão	1.95	2.01	4.83

Fonte: https://www.english-corpora.org/coca/

II. Terminologias

No corpus de tradução humana-oficial, um dos itens lexicais foi o adjetivo "*executive*". Há duas colocações muito comuns na tradução da área financeira com o adjetivo "*executive*": 1) "*executive officer*" (em português, diretor) e 2) "*executive board*" (em português, diretoria).

Ambas são traduções especificas para um contexto específico, exigindo do tradutor não só domínio do inglês financeiro, mas também experiência linguística para analisar esse contexto. O termo "diretor" que pede necessariamente a tradução como "*officer*" ou "*executive officer*" é ligado à diretoria ("*executive board*"), órgão colegiado executivo de administração, responsável pela gestão e representação da empresa. Esse órgão diretivo se faz mais presente em empresas com sistemas de governança corporativa, como as companhias de capital aberto.

Mas por qual motivo não traduzir "diretor" pela tradução 'mais óbvia', que seria "*director*"? Conforme as regulamentações aplicadas as empresas de capital aberto, as diretorias se reportam diretamente ao conselho de administração, que se reporta diretamente às assembleias gerais de acionistas. Na área financeira, membros do conselho de administração ou conselheiros da empresa são tipicamente denominados de "*directors*".

Traduzir "diretor" (sendo este membro da diretoria) como "*director*" (sendo este membro do conselho de administração), seria contextualmente incorreto, apesar de gramaticalmente correto. Tal tradução poderia induzir ao erro o público-alvo de tal tradução, inclusive acionistas da empresa ou até mesmo possíveis futuros investidores. Considerando que os textos disponibilizados pela empresa de capital aberto visam informar e reportar as condições da empresa para o público, tal erro levaria a um processo de desinformação, podendo até mesmo fazer com que a empresa sofra sanções de órgãos regulamentadores.

As terminologias em inglês americano são definidas com base nas regulamentações da *Securities and Exchange Commission* (Comissão de Valores Mobiliários dos Estados Unidos). A comissão estabelece que, para o Formulário D (relatório de prestação de contas para os investidores), as seguintes terminologias e definições devem ser observadas:

> "***Director***" *means any director of a corporation or any person performing similar functions with respect to any organization whether incorporated or unincorporated*. [...].

> (SEC, 2008).
>
> **"*Executive officer*"** *means the president, any vice president in charge of a principal business unit, division or function (such as sales, administration or finance), any other officer who performs a policy making function, or any other person who performs similar policy making functions for the issuer. Executive officers of subsidiaries may be deemed executive officers of the issuer if they perform such policy making functions for the issuer*[10].
>
> (SEC, 2008).

As terminologias em português são definidas pela Comissão de Valores Mobiliários. No documento "Recomendações da CVM sobre Governança Corporativa" de junho de 2002, disponibilizado no site da Comissão, encontramos as seguintes terminologias e definições.

> O conselho de administração deve atuar de forma a proteger o patrimônio da companhia, perseguir a consecução de seu objeto social e orientar a diretoria a fim de maximizar o retorno do investimento [...]. O conselho de administração deve ter de cinco a nove membros [...]. O mandato de todos os conselheiros deve ser unificado [...]. Os cargos de presidente do conselho de administração e presidente da diretoria (executivo principal) devem ser exercidos por pessoas diferentes. O conselho de administração fiscaliza a gestão dos diretores.
>
> (CVM, 2002).

Além disso, uma empresa de capital aberto tem dois tipos de presidentes. O presidente do Conselho de Administração (em inglês, "*Chairperson*" ou "*President*") e o presidente da Diretoria (em inglês, "*Chief Executive Officer*"), os dois com funções, responsabilidades e hierarquias completamente diferentes dentro da empresa. Tal diferenciação na arte da tradução atualmente requer um profissional linguístico experiente para analisar esse contexto e fazer as escolhas mais pertinentes, o que foi identificado pelo algoritmo J48.

4.2 Análise multidimensional funcional com Biber Tagger

[10] Tradução da autora: "Conselheiro" significa qualquer diretor de uma corporação ou qualquer pessoa que executa funções similares em relação a qualquer organização, seja ela pública ou privada. [...]. (SEC, 2008).
"Diretor" significa o presidente, qualquer vice-presidente responsável por uma unidade de negócios, divisão ou função fundamental (como vendas, administração ou finanças), qualquer outro diretor responsável por criar políticas, ou qualquer outro cargo similar responsável por criar políticas para a emissora. Diretores executivos das subsidiárias podem ser considerados diretores executivos da emissora, caso sejam responsáveis por criar políticas para a emissora. (SEC, 2008).

A análise multidimensional funcional gerou um resultado com baixo tamanho de efeito, representado pelo R-square (R^2), que foi de menos de 1% entre a tradução humana e a tradução automática. Esses resultados mostram que não há conjuntos de características gramaticais capazes de diferenciar os subcorpora estudados, pertencentes a um mesmo registro; sendo assim, essa análise efetuada não possibilitou a identificação de cada subcorpus, inferindo que são fracamente diferenciados pelas variáveis dimensionais. Em outras palavras, a tradução automática emula a tradução humana com bastante acurácia em termos das dimensões de variação do inglês.

Em seguida, decidimos treinar o classificador Weka para tentar classificar os textos. Os escores médios das cinco dimensões identificadas em Biber (1988) foram a base para serem calculados sobre os algoritmos J48 e Random Forest para quatro comparações diferentes, na plataforma SAS OnDemand for Academics:

1) Corpus Comparável vs. Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial;
2) Corpus Comparável vs. Corpus de Tradução Automática;
3) Corpus Comparável vs. Corpus de Tradução Humana-Oficial;
4) Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial.

Cada comparação gerou resultados diferentes e significativos para a atual pesquisa. Abaixo detalhamos cada uma dessas análises e possíveis conclusões.

4.2.1 Dimensões de variação e Algoritmo J48

I. Corpus Comparável vs. Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial

Ao utilizar os escores médios das cinco dimensões do inglês (BIBER, 1988) com o algoritmo J48 do Weka para os três corpora objetos de estudo desta pesquisa, o modelo probabilístico novamente dividiu as 2.716 amostras em dois grupos. O primeiro grupo, compreendendo 70% dos corpora (1.901 amostras), foi utilizado para treinamento do modelo. O segundo grupo, totalizando 815 amostras (30% dos corpora) foi utilizado como corpus de teste, para avaliar a precisão do modelo.

Os resultados mostraram que 815 amostras (51,53%) foram classificadas corretamente como tradução automática ou tradução humana ou texto provavelmente escrito por falante nativo. As classificações incorretas perfizeram 48,47% do total das amostras testadas.

Tabela 13 - Visão geral dos resultados do Biber Tagger e Algoritmo J48 (CS[11] vs. GTPS[12] vs. HCPS[13])

Amostras Classificadas Corretamente	420	51,53%
Amostras Classificadas Incorretamente	395	48.47%
Número Total de Amostras	815	100 %

Fonte: Autoria própria.

A tabela abaixo detalha a divisão do índice de acerto entre os três corpora utilizados nesta pesquisa. O corpus de tradução humana-oficial ficou muito próxima da média geral de acerto, no entanto, o corpus comparável e o corpus de tradução automática apresentaram resultados discrepantes.

O índice de acerto em relação ao corpus comparável foi bem acima da média geral (+22,17%), o que pode indicar maior precisão na identificação desse tipo de corpus ao utilizar as dimensões de Biber (1988) e o algoritmo J48 do Weka e também indicar que os marcadores lexicais de textos do corpus comparável foram mais pronunciados do que os marcadores lexicais das traduções.

Outro índice que chamou a atenção foi o do corpus de tradução automática, bem abaixo da média (-23,41%). Ao analisar estes resultados, pode-se inferir que a análise multidimensional linguística foi menos eficaz para esse tipo de corpus e/ou que há uma justaposição dos marcadores dimensionais do corpus de tradução automática e tradução humana.

Tabela 14 – Matriz de confusão

	Classificação inicial		
Classificação por máquina	Corpus comparável	Tradução automática	Tradução humana
Corpus Comparável	213	38	38
Tradução automática	60	72	124

[11] CS: *Comparable Subcorpora* (Corpus comparável com textos provavelmente escritos por falantes nativos da língua inglesa)
[12] GTPS: *Google Translate Parallel Subcorpus* (Corpus paralelo de tradução automática)
[13] HCPS: *Human Certified Parallel Subcorpus* (Corpus paralelo de tradução humana-oficial)

Tradução humana	82	53	135
Índice de Acerto	73,70%	28,12%	50,00%

Fonte: Autoria própria.

II. Corpus Comparável vs. Corpus de Tradução Automática

Ao comparar o corpus comparável com o corpus de tradução automática, por meio das cinco dimensões identificadas no estudo de Biber (1988) geradas no SAS OnDemand for Academics com o Biber Tagger, o algoritmo J48 analisou um total de 1.835 amostras, sendo 1.589 utilizadas pela aprendizagem por máquina para gerar a fórmula preditiva e as outras 550 amostras efetivamente testadas pelo algoritmo.

Dentre as amostras testadas, 398 foram classificadas corretamente e 152 classificadas incorretamente, sendo assim, o índice de acerto deste modelo probabilístico foi de 72,36% e o índice de erro foi de 27,63%. Apesar de o índice ser ligeiramente menor quando comparado ao algoritmo J48 sem as dimensões da análise multidimensional, o índice de acerto ainda assim é relativamente alto, podendo indicar que há diferenças significativas entre o corpus comparável e o corpus de tradução automática.

Tabela 15 - Visão geral dos resultados das dimensões de Biber (1988) e Algoritmo J48 (CS vs. GTPS)

Amostras Classificadas Corretamente	398	72,36%
Amostras Classificadas Incorretamente	152	27,64%
Número Total de Amostras	550	100 %

Fonte: Autoria própria.

Ao comparar os índices de acerto específicos de cada corpus (tabela abaixo), há uma ligeira diferença de menos de 10% entre os dois corpora. Quando comparados a média geral, os índices de acerto para cada corpus podem ser considerados dentro do desvio padrão esperado para essa análise.

Tabela 16 – Matriz de confusão

	Classificação inicial

Classificação por máquina	Corpus comparável	Tradução automática
Corpus comparável	219	66
Tradução automática	86	179
Índice de Acerto	76,84%	67,55%

Fonte: Autoria própria.

III. Corpus Comparável vs. Corpus de Tradução Humana-Oficial

O corpus comparável e corpus de tradução humana-oficial, considerando as cinco dimensões de Biber (1988), também foram comparados com algoritmo J48. O número de amostras total foi o mesmo da análise acima, assim como o total de amostras usadas como base e o total de amostras testadas também foram idênticos à análise anterior.

O índice de acerto também foi muito similar, chegando a 70,55%. No total, 388 amostras foram classificadas corretamente (10 a menos que na análise anterior) e 162 amostras foram classificadas incorretamente (10 a mais que na análise anterior). Isso indica que entre o corpus comparável e o corpus de tradução humana também há diferenças lexicais significativas.

Tabela 17 - Visão geral dos resultados das dimensões de Biber (1988) e Algoritmo J48 (CS vs. HCPS)

Amostras Classificadas Corretamente	388	70,55%
Amostras Classificadas Incorretamente	162	29,45%
Número Total de Amostras	550	100 %

Fonte: Autoria própria.

Novamente, ao analisar individualmente o corpus comparável e o corpus de tradução humana (tabela abaixo), observam-se índices de acerto bem diferentes da média. O algoritmo combinado às dimensões de Biber (1988) se mostra muito mais eficaz na classificação de textos do corpus comparável do que nas traduções humanas-oficiais.

Tabela 18 – Matriz de confusão

	Classificação inicial	
Classificação por máquina	Corpus comparável	Tradução humana
Corpus comparável	233	52
Tradução humana	110	155
Índice de Acerto	81,75%	58,49%

Fonte: Autoria própria.

É interessante pontuar que ambos os corpora de tradução, quando comparados ao corpus de textos provavelmente escritos por falantes nativos do inglês, demonstram diferenças lexicais importantes. Isso indica que há disparidades (conforme item V abaixo) entre textos escritos diretamente em inglês e traduções do português para o inglês. Além disso, é possível inferir, considerando que os índices de acerto são muito próximos, que há pouca diferença entre o corpus de tradução automática e humana-oficial, em termos lexicais, quando comparados ao corpus comparável.

IV. Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial

Ao combinar as cinco dimensões obtidas por meio da análise multidimensional de Biber (1988) com a análise de corpus via Weka com o algoritmo J48, o índice de acerto do modelo probabilístico cai de 83,26% – análise somente com o Weka sem as dimensões de Biber (1988) – para 50,85%, com o índice de erros em 49,15%.

Tabela 19 - Visão geral dos resultados do Biber Tagger e Algoritmo J48 (GTPS vs. HCPS)

Amostras Classificadas Corretamente	269	50,85%
Amostras Classificadas Incorretamente	260	49,15%
Número Total de Amostras	529	100 %

Fonte: Autoria própria.

Tivemos um total de 1.762 amostras nesta análise, sendo 70% das amostras utilizadas para aprendizagem por máquina (ou 'treino') e 30% restantes efetivamente testadas pelo método. Dentre as amostras testadas, 269 foram classificadas corretamente (170 a menos quando

comparada com a análise puramente via Weka, seguindo os mesmos critérios) e 260 foram classificadas incorretamente.

Tabela 20 – Matriz de confusão

	Classificação inicial	
Classificação por máquina	Tradução automática	Tradução humana
Tradução automática	17	256
Tradução humana	4	252
Índice de Acerto	6,23%	98,44%

Fonte: Autoria própria.

Apesar do índice médio de acerto entre os dois corpora ter sido relativamente baixo, ao analisar os índices separadamente os resultados foram impressionantes. O índice de acerto do corpus de tradução humana-oficial foi de surpreendentes 98,44%. Já o índice de acerto do corpus de tradução automático foi extremamente abaixo do esperado, chegando a 6,23%.

É possível inferir duas características interessantes dessa análise: 1) Os indicadores lexicais das traduções humanas, considerando as dimensões de Biber (1988), foram extremamente precisos e eficazes; e 2) Os indicadores lexicais das traduções humanas se fizeram muito presentes nas traduções automáticas, ou seja, a aprendizagem por máquina foi incapaz de identificar as traduções automáticas devido a significativa presença de marcadores lexicais presentes em traduções humanas.

Portanto, observa-se que, nesta análise especificamente, o algoritmo identificou com precisão os indicadores lexicais de tradução humana e as traduções automáticas foram tão parecidas com as traduções humanas, que os indicadores de tradução automática foram ineficazes em separar uma tradução da outra e classificar corretamente as traduções automáticas.

Sendo assim, tais números indicam que a análise via Weka com o algoritmo J48 combinada com a análise via as dimensões de Biber (1988) foi menos eficaz e equilibrada que a análise que não empregava as dimensões de variação.

V. Exemplos de disparidades terminológicas

Comparando três documentos de duas companhias de capital aberto do setor de locação de veículos, uma sediada no Brasil e outra sediada nos Estados Unidos, encontramos as disparidades mencionadas no item III. Ambos os documentos tratam do mesmo tema, divulgando um programa de recompra de ações autorizado pelo Conselho de Administração. O texto em português da companhia brasileira estava também disponível em inglês no site de relações com investidores. Além disso, o texto em português foi traduzido de forma automática no MemoQ.

A partir destes três textos, notamos as disparidades terminológicas abaixo. É possível notar que há forte similaridade entre os textos traduzidos, tanto por humanos quanto pela máquina. Já o texto provavelmente escrito por um falante nativo da língua inglesa traz diferenças significativas em relação à terminologia.

Tabela 21 – Exemplos de diferenças terminológicas entre textos da área financeira sobre o mesmo tema e de companhias similares com a mesma área de atuação

ENG-TRAD	ENG-TM	ENG-ORIG
on this date	*on this date*	*today*
informs	*Informs*	*announced*
approved	*approved*	*authorized*
buyback program of shares	*program for the repurchase of shares*	*share repurchase program*
shares outstanding in the market	*shares outstanding in the market*	*outstanding common stock*
The program aims	*The purpose of the Program is*	*The repurchase program allows*
in order to comply with	*in order to meet the*	*in accordance with*

Fonte: Autoria Própria.

Os textos acima não fizeram parte do corpus analisado e servem puramente para propósitos exemplificativos. Ainda assim, vale ressaltar que os textos similares aos selecionados para compor essa tabela foram abundantes no corpus objeto desta pesquisa e, portanto, pode-se supor que os exemplos destes textos, apesar de não fazerem parte do corpus, também se aplicam aos corpora estudados.

4.2.2 Dimensões de Biber (1988) e Algoritmo Random Forest

I. Corpus Comparável vs. Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial

No processamento de dados via Algoritmo Random Forest utilizando as cinco dimensões geradas a partir das dimensões de Biber (1988), considerando todos os corpora desta pesquisa (ou seja, corpus de textos em inglês provavelmente escritos por falante nativo, corpus de traduções automáticas e corpus de traduções oficiais disponibilizadas pelas companhias de capital aberto, nesse estudo consideradas traduções realizadas por profissionais de serviços linguísticos), obtivemos os resultados abaixo.

Totalizamos 2.176 amostras, sendo que 30% (815) destas amostras foram testadas pelo método probabilístico desenvolvido a partir da aprendizagem por máquina com o processamento de 70% das amostras como treino. Deste total, 446 amostras foram classificadas corretamente (54,72%) e 369 amostras foram classificadas incorretamente (45,28%).

Tabela 22 - Visão geral dos resultados das dimensões de Biber (1988) e Algoritmo Random Forest (CS vs. GTPS vs. HCPS)

Amostras Classificadas Corretamente	446	54,72%
Amostras Classificadas Incorretamente	369	45,28%
Número Total de Amostras	815	100 %

Fonte: Autoria própria.

Na tabela abaixo é possível observar que o algoritmo Random Forest, ao processar as cinco dimensões de Biber (1988), teve um bom desempenho (índice de acerto de 75,43%) em relação ao corpus comparável. A tradução automática e a tradução humana tiveram mais erros que acertos (índice de acerto de 42,19% e 44,44%, respectivamente).

Tabela 23 – Matriz de confusão

	Classificação inicial		
Classificação por máquina	Corpus comparável	Tradução automática	Tradução humana
Corpus comparável	218	40	31
Tradução automática	50	108	98
Tradução humana	60	90	120
Índice de Acerto	75,43%	42,19%	44,44%

Fonte: Autoria própria.

Mais uma vez, de forma consistente com as análises anteriores, o algoritmo se provou mais preciso e eficaz quando comparando o corpus comparável com o corpus de tradução e menos eficaz ao comparar o corpus de tradução automática com o corpus de tradução humana. Isso indica, novamente, uma possível maior discrepância entre o corpus comparável e o corpus de tradução e uma maior similaridade entre os corpora de tradução, independentemente de as traduções serem automáticas ou humanas-oficiais.

II. Corpus Comparável vs. Corpus de Tradução Automática

Ao comparar o corpus comparável com o corpus de tradução automática com as cinco dimensões de Biber (1988) e usando o algoritmo *Random Forest*, obtivemos os seguintes resultados: 76% das amostras foram classificadas corretamente e 24% das amostras foram classificadas incorretamente. Foram 550 amostras testadas (30%) e 1.285 amostras usadas para treino (70%), totalizando 1.835 amostras.

Tabela 24 - Visão geral dos resultados das dimensões de Biber (1988) e Algoritmo Random Forest (CS vs. GTPS)

Amostras Classificadas Corretamente	418	76,00%
Amostras Classificadas Incorretamente	132	24,00%
Número Total de Amostras	550	100 %

Fonte: Autoria própria.

Na tabela acima, o índice de acerto para amostras do corpus comparável foi maior que o índice das amostras do corpus de tradução automática, conforme a tabela abaixo. Esse dado é congruente com os resultados da análise utilizando as dimensões de Biber (1988).

Tabela 25 - Matriz de confusão

	Classificação inicial	
Classificação por máquina	Corpus comparável	Tradução automática
Corpus comparável	238	47

Tradução automática	85	180
Índice de Acerto	83,51%	67,94%

Fonte: Autoria própria.

III. Corpus Comparável vs. Corpus de Tradução Humana-Oficial

Na análise do corpus comparável combinado ao corpus de tradução humana-oficial, foram processadas 1.835 amostras, sendo que 70% deste total foi utilizado para treinar o método probabilístico e 30% foi efetivamente testado pela ferramenta com o algoritmo Random Forest.

Sendo assim, no total, 550 amostras foram testadas, com os seguintes resultados: 414 amostras foram classificadas corretamente, perfazendo um índice de acerto de 75,27%, e 136 amostras foram classificadas incorretamente, com um índice de acerto de 24,73%.

Tabela 26 - Visão geral dos resultados do Biber Tagger e Algoritmo Random Forest (CS vs. HCPS)

Amostras Classificadas Corretamente	414	75,27%
Amostras Classificadas Incorretamente	136	24,73%
Número Total de Amostras	550	100 %

Fonte: Autoria própria.

Na tabela abaixo, os dois corpora foram separados, visando uma maior compreensão dos dados gerados nesta análise. O índice de acerto do corpus comparável foi ligeiramente maior que o índice de acerto do corpus de tradução humana, mas dentro do desvio padrão médio esperado.

Tabela 27 - Matriz de confusão

	Classificação inicial	
Classificação por máquina	Corpus comparável	Tradução humana
Corpus comparável	234	51
Tradução humana	85	180
Índice de Acerto	82,11%	67,92%

Fonte: Autoria própria.

IV. Corpus de Tradução Automática vs. Corpus de Tradução Humana-Oficial

Por fim, na análise do corpus de tradução automática comparado ao corpus de tradução humana os dois corpora totalizaram 1.762 amostras, divididas em cinco dimensões geradas pelo Biber Tagger no SAS OnDemand for Academics. Conforme as demais análises, utilizamos 70% das amostras como treino e 30% (529 amostras) foram testadas. Na tabela abaixo, é possível observar o número de amostras classificadas corretamente e incorretamente e seus respectivos índices de acerto.

Tabela 28 - Visão geral dos resultados das dimensões de Biber (1988) e Algoritmo Random Forest (GTPS vs. HCPS)

Amostras Classificadas Corretamente	271	51,23%
Amostras Classificadas Incorretamente	258	48,77%
Número Total de Amostras	529	100 %

Fonte: Autoria própria.

Abaixo, detalhamos os índices de acerto com base no corpus individual testado. O índice de acerto de ambos os corpora foi muito próximos um do outro, com uma diferença de cerca de 6% entre o corpus de tradução automática e o corpus de tradução humana-oficial.

Tabela 29 – Matriz de confusão

	Classificação inicial	
Classificação por máquina	Tradução automática	Tradução humana
Tradução automática	148	125
Tradução humana	133	123
Índice de Acerto	54,21%	48,05%

Fonte: Autoria própria.

4.3 Visão geral

Na tabela abaixo, incluímos todos os índices vistos acima, gerais e específicos, para aprofundar a discussão e aprofundar a análise dos resultados.

Tabela 30 – Visão geral do índice de acerto de todos os resultados com Algoritmo Random Forest, Algoritmo J48 e dimensões de Biber (1988)

Weka	**CS**	**GTPS**	**HCPS**	**Média**
Random Forest	-	84,38%	83,88%	82,12%
J48 (70%/30%)	-	87,91%	77,73%	82,99%
J48 (60%/40%)	-	84,28%	82,25%	83,26%
Biber (1988) + Weka	**CS**	**GTPS**	**HCPS**	**Média**
J48 (CS x GTPS x HCPS)	73,70%	28,12%	50,00%	51,53%
J48 (CS x GTPS)	76,84%	67,55%	-	72,36%
J48 (CS x HCPS)	81,75%	-	58,49%	70,55%
J48 (GTPS x HCPS)	-	6,23%	98,44%	50,85%
RF (CS x GTPS x HCPS)	75,43%	42,19%	44,44%	54,72%
RF (CS x GTPS)	83,51%	67,94%	-	76,00%
RF (CS x HCPS)	82,11%	-	67,92%	75,27%
RF (GTPS x HCPS)	-	54,21%	48,05%	51,23%

Fonte: Autoria própria.

Chave:

RF = Random Forest

CS = Corpus Comparável

GTPS = Corpus de Tradução Automática

HCPS = Corpus de Tradução Humana-Oficial

O processamento de dados via Weka, tanto com o algoritmo Random Forest quanto com o algoritmo J48, foi consistente, com variação entre 83,26%, sendo o índice de acerto mais alto, e 82,12%, sendo o menor índice de acerto, ou seja, uma variação de 1,14% para mais ou para menos.

Conforme observado acima, há marcadores lexicais importantes ao se comparar as traduções automáticas e as traduções humanas, possibilitando o uso da ferramenta para identificar estatisticamente, por meio deste modelo preditivo e com um índice de acerto relativamente alto, se a tradução testada é uma tradução puramente por máquina ou uma tradução realizada por um tradutor ou prestador de serviços linguísticos.

O processamento de dados via Weka com as dimensões de Biber (1988), que analisam os corpora do ponto de vista gramatical, gerou dados interessantes, apesar do índice de acerto geral relativamente baixo na análise do conjunto total de corpora (CS x GTPS x HCPS) e na análise do corpora de traduções (GTPS x HCPS).

Das oito análises realizadas com os corpora acima, o resultado variou entre 54,72% e 50,85%, uma variação de 3,89% para mais ou para menos. Pode-se inferir que, a partir destes resultados, há poucos marcadores gramaticais que diferenciam os corpora de tradução, tendo em vista que, sempre que comparados, os índices de acerto foram medianos.

No entanto, é importante mencionar que, nestas análises, o índice de acerto do corpus comparável, quando presente, foi muito acima da média geral, indicando que o corpus comparável apresenta mais diferenças gramáticas que o corpus de tradução, independentemente do tipo de tradução. Na análise com o algoritmo J48, o índice de acerto do corpus comparável foi de 73,70% (22,17% acima da média), e na análise com o Random Forest, o índice de acerto foi de 75,43% (20,71% acima da média).

Outro dado que merece destaque foi observado na análise com o algoritmo J48 e as dimensões de Biber (1988) acerca do corpus de tradução automática e tradução humana. O índice de acerto da tradução humana foi de 98,44%, enquanto a do corpus de tradução automática foi de 6,23%. Este dado em específico requer mais análises, com outras porcentagens de treino e, talvez, até mesmo outros métodos de processamento de dados, pois foi completamente inconsistente com as demais análises.

Nas análises comparativas entre o corpus comparável e o corpus de tradução automática ou humana, os índices de acerto foram melhores, variando entre 76,00% e 70,55% (variação de 5,45%). Na análise discriminada, o corpus comparável obteve um melhor desempenho em comparação com o corpus de tradução. Isso indica que há marcadores gramaticais que diferenciam e distanciam o corpus de textos provavelmente escritos por falantes nativos da língua inglesa e o corpus de tradução com textos partindo do português para o inglês.

A presença desses marcadores gramaticais entre o corpus comparável e o corpus de tradução, corrobora algo há muito tempo estudado por acadêmicos da área de estudos tradutórios e teoria da tradução, a inexistência da equivalência entre o texto original e o texto traduzido. Uma tradução nunca será perfeita, devido à impossibilidade de transpor toda a carga cultural de um idioma para o outro.

Conforme afirmado na edição especial "Tradução, desconstrução e pós-modernidade":

> O texto traduzido é "outro" texto, que mantém outro tipo de relações entre os elementos, exatamente porque as coerções impostas pelas línguas levam a diferentes possibilidades de contextualizações, de remissões, de encadeamentos, de atribuição de valores entre os elementos. Essas concepções poderiam levar a se pensar que a tradução é totalmente impossível. No entanto, o que é impossível não é a tradução, mas a noção de tradução de que se parte para pensar nessa impossibilidade: uma concepção que espera que a tradução repita o texto original, que seja seu equivalente, que reproduza seus valores.
> (RODRIGUES, 2001, p. 95).

Por fim, pode-se inferir que há pouca diferença gramatical entre a tradução automática e a tradução humana. Sendo este o caso, apesar de lexicalmente os corpora de tradução apresentarem diferenças significativas, gramaticalmente são pouco distintos.

5 Considerações finais

Esta pesquisa teve dois principais objetivos: 1) investigar a existência de traços linguísticos entre traduções humanas e traduções automáticas e 2) investigar a existência de traços linguísticos diferenciando textos traduzidos de textos provavelmente escritos por falantes nativos. Para tais investigações, selecionamos registros da área financeira, especificamente do mercado de ações, com empresas de capital aberto. Essa seleção foi feita devido à complexidade terminológica, às características marcantes de textos em inglês e português e às diferentes regulamentações para empresas com ações negociadas no mercado brasileiro e no mercado norte-americano.

Visando diversificar o corpus entre diferentes subsegmentos inseridos nesse registro, selecionamos dez setores de atuação diferentes e, portanto, dez empresas brasileiras com ações negociadas no mercado americano e no mercado brasileiro (perfazendo os corpora de tradução automática e de tradução humana) e dez empresas americanas com ações negociadas no mercado americano.

As dimensões de variação de Biber (1988) se mostram pouco preditivas da diferença entre os tipos de tradução, com R^2 menor que 1%. Esse resultado demonstra que praticamente não há diferença entre os subcorpora analisados, do ponto de vista gramatical.

Já a análise lexical, por meio dos algoritmos J48 e Random Forest do Weka, conseguiu distinguir os tipos de tradução com maior eficiência, chegando a um índice de acerto de mais de 83% com o algoritmo J48 e de mais de 82% com o algoritmo Random Forest em relação aos corpora paralelos (tradução automática e tradução humana). Ou seja, do ponto de vista lexical, há uma diferença significativa marcante entre a tradução humana e a com máquina. Por fim, ao combinar os dois métodos de análise, foi obtido alto índice de acerto, variando entre 83% e 73%.

Sendo assim, é possível inferir que, do ponto de vista lexical, há variações linguísticas que permitem diferenciar e, de forma probabilística e estatística, prever com certa precisão se um texto trata-se de uma tradução ou de um texto possivelmente produzido em inglês por nativos do idioma.

Retomando as perguntas iniciais da pesquisa é possível chegar as seguintes conclusões:

1. Há traços linguísticos que podem diferenciar de forma probabilística a tradução humana da tradução automática? A resposta é afirmativa. Utilizando o léxico, é possível obter um índice de acerto probabilístico superior a 83% para distinguir entre tradução humana e

tradução automática. Já do ponto de vista gramatical, a análise sugere que não há diferença significativa entre a tradução humana e automática.

2. Há traços linguísticos que podem diferenciar de forma probabilística e estatística uma tradução para o inglês (seja ela automática ou humana) de um texto produzido originalmente em inglês? A resposta é afirmativa. Foi possível obter um índice de acerto superior a 76% para distinguir entre a tradução automática e textos provavelmente escritos por falantes nativos e superior a 75% para distinguir entre a tradução humana e textos provavelmente escritos por falantes nativos. Do ponto de vista gramatical, no entanto, não foi possível detectar diferenças entre as categorias de texto.

Com base nesses resultados, há inúmeras possibilidades de continuação para essa pesquisa. No ponto de vista da autora, dois potenciais são mais evidentes com base nos números e nas aplicabilidades materiais da pesquisa. O primeiro potencial é o uso da metodologia para criar um sistema probabilístico de identificação de tradução automática em textos da área financeira, com índices de acerto altos do ponto de vista estatístico. O segundo potencial é o uso da metodologia para aprimorar as traduções financeiras por meio da formação de tradutores especializados ao utilizar conteúdos específicos da área desenvolvidos com base nos resultados e assim buscar aproximar a tradução dos textos provavelmente escritos por falantes nativos.

Além das possibilidades acima, há diversas outras possibilidades, tais como uma análise qualitativa com avaliações de qualidade das traduções com amostras randomizadas para complementar esse estudo inicial que levou em conta somente analises quantitativas; uma investigação de porcentagem de aproveitamento da tradução automática partindo do pressuposto de que a tradução humana foi uma pós-edição de tradução automática; ou até mesmo análises quantitativas e qualitativas separadamente de cada um dos setores de atuação e dos tipos de textos selecionados e coletados no corpus em questão.

6 Referências bibliográficas

ABDALLAH, K. **Translators in industry: an investigation of translation practices in the private sector.** The Interpreter and Translator Trainer, v. 6, n. 2, p. 167-184, 2012.

BAKER M. **Corpora in translation studies: An overview and some suggestions for future research.** Target, v. 8, n. 2, p. 253-280, 1995.

BAKER, M. Corpus linguistics and translation studies: Implications and applications. In: BAKER, M., FRANCIS, G. & TOGNINI-BONELLI, E. (Eds.). **Text and Technology**: **In Honour of John Sinclair.** John Benjamins, p. 233-250, 1993.

BAKER, M. (Ed.). **Routledge Encyclopedia of Translation Studies**. Abingdon, Oxfordshire: Routledge, 1998.

BAKER, M. **Towards a methodology for investigating the style of a literary translator**. Target, v. 12, n. 2, p. 241-266, 2000.

BAKER, P. **Sociolinguistics and corpus linguistics**. Edinburgh: Edinburgh University Press, 2010.

BERBER SARDINHA, T. **Análise multidimensional.** DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, v. 16, n. 1, p. 99-127, 2000.

BERBER SARDINHA, T. **Linguística de corpus**. Barueri: Editora Manole, 2004.

BERBER SARDINHA, T. **A historical characterisation of American and Brazilian cultures based on lexical representations.** Corpora, v. 15, n. 2, p. 183-212, 2020.

BERBER SARDINHA, T. & VEIRANO PINTO, M. (Eds.). **Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber** (Vol. 60). Amsterdam: John Benjamins Publishing Company, 2014.

BERBER SARDINHA, T. & VEIRANO PINTO, M. **Multi-dimensional analysis: Research methods and current issues.** London: Bloomsbury Academic, 2019.

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988.

BIBER, D. **Dimensions of register variation**: **A cross-linguistic comparison.** Cambridge: Cambridge University Press, 1995.

BIBER, D., CONRAD, S. & REPPEN, R. **Corpus linguistics: Investigating language structure and use.** Cambridge: Cambridge University Press, 1998.

BIBER, D. & CONRAD, S. **Register, genre, and style.** Cambridge: Cambridge University Press, 2019.

BOWKER, L. **Computer-Aided Translation Technology: A Practical Introduction.** Ontario: University of Ottawa Press, 2002.

BRITISH STANDARDS INSTITUTION. **ISO 17100: 2015 Translation Services– Requirements for Translation Services.** 2015.

BRITISH STANDARDS INSTITUTION. **ISO 18587:2017 Translation Services– Post-editing of machine translation output — Requirements.** 2017.

COMISSÃO DE VALORES MOBILIÁRIOS. **Recomendações da CVM sobre governança corporativa.** 2002. Recuperado em 06 de fevereiro de 2023. https://conteudo.cvm.gov.br/export/sites/cvm/decisoes/anexos/0001/3935.pdf.

DRUGAN, J. **Quality in professional translation: assessment and improvement**. London: Bloomsbury Publishing, 2013.

EGBERT, J., BIBER, D. & GRAY, B. **Designing and evaluating language corpora: A practical framework for corpus representativeness.** Cambridge: Cambridge University Press, 2022.

FRANCIS, W. N. & KUCERA, H. **Brown corpus manual.** Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

FUKARI, A. & WOLF, M. **Constructing a sociology of translation**. Amsterdam: John Benjamins Publishing, 2007.

GARCÍA, I. S. **A procedural approach to the translation of financial statements**. Meta: Journal des traducteurs /Meta: Translators' Journal, v. 60, n. 1, p. 3-23, 2015.

GASPARI, F., ALMAGHOUT, H. & DOHERTY, S. **A study on the use of machine translation in corporate settings: Is the user the weakest link**? In: Proceedings of the Workshop on Humans and Computer-Assisted Translation (HaCAT), p. 53-61, 2014.

GENTZLER, E. **Translation and identity in the Americas: New directions in translation theory**. Abingdon, Oxfordshire: Routledge, 2008

HOLMES G., DONKIN A. & WITTEN I. H. **WEKA: a machine learning workbench.** Proceedings of ANZIIS '94 - Australian New Zealand Intelligent Information Systems Conference, Brisbane, QLD, Australia, p. 357-361, 1994.

HOUSE, J. **Text and context in translation**. Journal of Pragmatics, Volume 38, Issue 3, p. 338-358, 2006.

HUTCHINS, W. J. Machine Translation: A Brief History. In: **Concise history of the language sciences: from the Sumerians to the cognitivists.** Edited by KOERNER E. F. K. & ASHER R. E. Oxford: Pergamon Press, p. 431-445, 1995.

HUTCHINS, W. J. & SOMERS, H. L. **An introduction to machine translation**. Cambridge: Academic Press, 1992.

JOHANSSON, S. **Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.** ICAME collection of English language corpora, v. 4. Oslo: University Department of English, 1978.

KENNY, D. **Corpus-based translation studies: A quantitative or qualitative development?**. Journal of Translation Studies v. 9, no. 1, p. 43-58, 2006.

KENNY, D. & DOHERTY, S. **Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators**. The Interpreter and Translator Trainer, v. 8, n. 2, p. 276-294, 2014.

KOEHN, P. **Europarl: A parallel corpus for statistical machine translation**. In: Proceedings of machine translation summit x: papers, p. 79-86, 2005.

LAVIOSA, S. **Core patterns of lexical use in a comparable corpus of English narrative prose**. Meta, v. 43, n. 4, p. 557-570, 1998.

LAVIOSA, S. **Corpus-based Translation Studies: Theory, Findings, Applications**. Amsterdam; New York: Rodopi, 2002.

MANNING, C. & SCHUTZE, H. **Foundations of Statistical Natural Language Processing.** Cambridge; Massachusetts; London; England: The MIT Press, 1999.

MICHEL, J.B. et.al. **Quantitative Analysis of Culture Using Millions of Digitized Books.** Science, v. 331, p. 176–182, 2011.

MCENERY, T. & WILSON, A. **Corpus Linguistics: An Introduction.** 2nd Edition, Edinburgh: Edinburgh University Press, 2001.

MCENERY, T. & HARDIE, A. **Corpus linguistics: Method, theory and practice**. Cambridge: Cambridge University Press, 2011.

MCENERY, T., XIAO, R. & TONO, Y. **Corpus-Based Language Studies: An Advanced Resource Book.** Abingdon, Oxfordshire: Routledge, 2006.

O'BRIEN, S. **Towards predicting post-editing productivity.** Machine Translation, v. 25, n. 3, p. 197-215, 2011.

O'BRIEN, S. **Towards a Dynamic Quality Evaluation Model for Translation**. Journal of Specialised Translation n. 17, p. 55-77, 2012.

O'HAGAN, M. & ASHWORTH, D. **Translation-mediated Communication in a Digital World: Facing the Challenges of Globalization and Localization**. Bristol, Blue Ridge Summit: Multilingual Matters, 2002.

OLOHAN, M. **Introducing Corpora in Translation Studies**. Abingdon, Oxfordshire: Routledge, 2004.

PIETRZAK, P. & KORNACKI, M. **Using CAT Tools in Freelance Translation: Insights from a Case Study**. Abingdon, Oxfordshire: Routledge, 2020.

PYM et al., **The Status of the Translation Profession in the European Union**. London: Anthem Press, 2012.

RESENDE, S. V. **Dimensões de variação do texto traduzido: uma abordagem multidimensional.** 2019.

RODRIGUES, C. C. **Tradução: a questão da equivalência**. ALFA: Revista de Linguística, São Paulo, v. 44, n. 1, p. 89-98 2001. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4281. Acesso em: 15 abr. 2023.

RENOUF, A., KEHOE, A. & BANERJEE, J. **WebCorp: an integrated system for web text search**. In: Corpus Linguistics and the Web, v. 59, p. 47-67, 2007.

SAS. **SAS OnDemand for Academics**. 2021. Recuperado em 15 de abril de 2023, de https://www.sas.com/pt_br/learn/on-demand-for-academics.html

SECURITIES AND EXCHANGE COMMISSION. **Final rule: plain English disclosure**. Release Nos. 33-7497; File No. S7-3-98. 2000.

STUBBS, M. **Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies**. Functions of Language, vol. 2:1, p. 23-55, 1995.

SVARTVIK, J. The **London–Lund corpus of spoken English: Description and research**. Lund: Lund University Press. 1990.

VENUTI, L. (Ed.). **The Translation Studies Reader.** Abingdon, Oxfordshire: Routledge, 2004.

WAY, A. Human and automatic translation: A symbiotic relationship. In: MOORKENS J., CASTILHO S., GASPARI F. & GASPARINI D. (Eds.). **Translation Quality Assessment: From Principles to Practice.** New York: Springer International Publishing, p. 27-41, 2018.

WU, Y. et al. **Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation**. Cornell University, arXiv:1609.08144v2, 2016.

ZANETTIN, F., BERNARDINI, S., & STEWART, D. **Corpora in Translator Education**. Abingdon, Oxfordshire: Routledge, 2003.