PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
FACULDADE DE FILOSOFIA, COMUNICAÇÃO, LETRAS E ARTES

André Carballido Dominguez

**Estrangeirismo na língua portuguesa: o uso de anglicismos na rede social do *Twitter* do
Brasil**

GRADUAÇÃO EM LETRAS: LÍNGUA INGLESA - TRADUÇÃO: INGLÊS/PORTUGUÊS

São Paulo

2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
FACULDADE DE FILOSOFIA, COMUNICAÇÃO, LETRAS E ARTES

André Carballido Dominguez

“Estrangeirismo na língua portuguesa: o uso de anglicismos na rede social do *Twitter* do Brasil”

Trabalho de Conclusão de Curso apresentado à Banca Examinadora da Pontifícia Universidade Católica, como exigência parcial para obtenção do título de Bacharelado em Letras: Língua Inglesa – Tradução Inglês/Português sob a orientação do Professor Doutor Tony Berber Sardinha.

São Paulo

2022

André Carballido Dominguez

“Estrangeirismo na língua portuguesa: o uso de anglicismos na rede social do *Twitter* do Brasil”

Trabalho de Conclusão de Curso apresentado à Banca Examinadora da Pontifícia Universidade Católica, como exigência parcial para obtenção do título de Bacharelado em Letras: Língua Inglesa – Tradução Inglês/Português.

Aprovado em: __/__/__

BANCA EXAMINADORA

Dr. Tony Berber Sardinha – PUCSP

Dra. Eliane Trindade – PUCSP

Dra. Maria Cláudia Nunes Delfino – FATEC

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, Flávia e José, por sempre acreditarem em mim, sempre apoiarem minhas escolhas e por dar suporte quando necessário. Seu amor e carinho foram fundamentais para minha vida como ser humano, vocês me prepararam para o mundo, devo tudo a vocês.

Agradeço a Beatriz, minha amiga e parceira, que esteve ao meu lado desde o começo e em todas as vezes que precisei durante essa jornada. Seu amor, incentivo, carinho, paciência e compreensão foram cruciais para chegar aonde cheguei e são a força que preciso para enfrentar o amanhã.

Agradeço aos meus amigos, que de diversas formas sempre me incentivaram e ajudaram a chegar neste momento e a me tornar a pessoa que sou hoje.

Agradeço ao meu orientador, Tony Berber Sardinha, não só pela orientação, mas pelo seu conhecimento único da área, sempre me dando confiança e apoio, crucial para o desenvolvimento deste trabalho.

Por fim, porém não menos importante, agradeço a todos que já passaram pela minha vida, que apesar de não estarem mais presentes, sempre estarão em meu coração.

RESUMO

Por ser uma língua mundial, o inglês é muito utilizado em muitos setores da sociedade, sendo incorporado em diferentes línguas de todo o mundo. A língua inglesa atingiu esse estatuto de língua mundial devido à dominação política, econômica e cultural dos países anglófonos, especialmente o Reino Unido e Estados Unidos. Ao mesmo tempo em que o inglês se torna uma língua tão influente, surgem movimentos de resistência à utilização do inglês em muitos países, como França e Brasil, entre outros. No caso do Brasil, questiona-se o emprego ‘exagerado’ e indiscriminado de ‘ estrangeirismos’, o que estaria colonizando e desvirtuando a língua portuguesa. Com a chegada da pandemia, percebe-se que o uso da língua inglesa na comunicação em língua portuguesa talvez tenha se acelerado. Seja na comunicação falada como nas redes sociais, tem-se a percepção de que a os brasileiros incorporam cada vez mais palavras inglesas no vocabulário. Embora haja essa percepção de que existe um uso crescente de palavras estrangeiras no português, o qual pode ter-se acelerado com a pandemia, não existem pesquisas que enfoquem diretamente a incidência do uso de palavras da língua inglesa no português brasileiro. Desta forma, este trabalho tem como objetivo analisar a incidência de palavras da língua inglesa em postagens da rede social Twitter de 2019 a 2022. Para isso, foi criado um corpus com os tweets (postagens) de 200 usuários variados, dentre eles 50 usuários governamentais, 50 usuários de pessoas consideradas influenciadores, 50 usuários de empresas de diferentes atuações e 50 usuários de pessoas ‘comuns’, do período de janeiro de 2019 a outubro de 2022. A base teórica da pesquisa consiste da Linguística de Corpus, uma área dos Estudos da Linguagem que se ocupa da coleta e análise de grandes quantidades de texto por meio de tecnologia computacional (BERBER SARDINHA, 2004). Além do corpus de tuítes criado especificamente para esta pesquisa, foram usados corpora de língua inglesa e língua portuguesa, bem como ferramentas de coleta em lote de postagens (*snsrape*), um script desenvolvido para o projeto pelo professor orientador e o pacote SAS de análise estatística. A partir disso, foi possível responder as seguintes perguntas: “Quais as médias de uso de palavras em inglês mais usadas no *Twitter* em português do Brasil?”; “Existe diferença na incidência de uso de palavras em inglês entre domínios/setores sociais e econômicos do *Twitter* do Brasil?”; “Existe diferença na incidência entre usuários em relação ao uso de palavras em inglês no *Twitter* em português do Brasil?”; e “Existe variação temporal no uso de palavras em inglês no *Twitter* do Brasil? Mais especificamente, existe um aumento ou diminuição no uso de palavras em inglês tendo como ponto de referência a pandemia de COVID-19?”. A análise permitiu

verificar o alcance do emprego da língua inglesa em postagens em português da rede social *Twitter*.

Palavras-chave: Estrangeirismo; anglicismo; linguística de corpus; *Twitter*; redes sociais; pandemia; COVID-19.

ABSTRACT

Being a world language, English is widely used in many sectors of society, in different languages around the world. English language has reached the current status of a world language due to the political, economic and cultural domination of English-speaking countries, especially the United Kingdom and the United States. As English vocabulary makes its way into more languages at a seemingly increasing pace, resistance against English hegemony grows in many countries, such as France and Brazil. In the case of Brazil, the ‘exaggerated’ and indiscriminate use of ‘foreignisms’, is commonly seen as a reflection of colonization, which would lead to the takeover of the local culture. With the COVID-19 pandemic, it is our impression that the use of the imported English vocabulary may have increased. Despite this growing perception, there is no empirical research looking at the incidence of the borrowing of English vocabulary in Brazilian Portuguese. Thus, this work aims to analyze the frequency of English words on the social network Twitter from 2019 to 2022. To meet this goal, a corpus was collected comprising tweets from 200 different users, broken down into groups or domains, namely 50 government users, 50 influencers, 50 corporations and 50 ‘ordinary’ people, from January 2019 to October 2022. The theory and method for the current project are based on Corpus Linguistics, a sub-discipline of Linguistics geared toward the collection and analysis of large amounts of text using computational technology (BERBER SARDINHA, 2004). In addition to the corpus of tweets created specifically for this project, other English and Portuguese language corpora were used to enable the (semi-)automatic identification of English borrowings in the tweets. The tools used in the analysis include the snsrape package for scraping the tweets, a script developed for the project by the supervisor, and the SAS package for statistical analysis. The research questions were as follows: “What is the mean frequency of English borrowings in Twitter posts written in Brazilian Portuguese?”; “Is there a difference of frequency in the use of English words across the different domains?”; “Is there a difference of frequency across the individual users?”; and “Is there diachronic variation in the use of English borrowings, by year, and between pre-pandemic and pandemic time frames?”. The results provided answers to these questions, which are presented in the report.

Keywords: Foreignism, anglicism, corpus linguistics, Twitter, social network, pandemic, COVID-19.

LISTA DE TABELAS

Tabela 1 — Design do corpus iWeb.....	20
Tabela 2 — Design do corpus CLIMA-CLIC	20
Tabela 3 — Os 10 vocábulos mais usados em inglês no domínio empresarial.....	23
Tabela 4 — Os 10 vocábulos mais usados em inglês no domínio dos famosos	24
Tabela 5 — Os 10 vocábulos mais usados em inglês no domínio do governo	24
Tabela 6 — Os 10 vocábulos mais usados em inglês no domínio de pessoas comuns.....	25
Tabela 7 — Média de uso de palavras inglesas por domínio	26
Tabela 8 — Postagens com palavras inglesas por domínio.....	26
Tabela 9 — Relação dos 20 usuários com maior média de uso de palavras em inglês.....	27
Tabela 10 — Média de uso de palavras inglesas em cada ano.....	28
Tabela 11 — Média de uso de palavras em inglês por trimestre.....	28

SUMÁRIO

1. INTRODUÇÃO	9
2. FUNDAMENTAÇÃO TEÓRICA	13
2.1. Linguística de corpus (LC)	13
2.2. Estrangeirismos.....	16
3. METODOLOGIA	18
3.1. Design e coleta do corpus de pesquisa.....	18
3.1.1. Critérios para seleção dos usuários.....	18
3.2. <i>Scraping</i>	19
3.3. Limpeza de corpus	20
3.4. Corpus de Referência.....	20
3.5. Indexação do corpus	21
3.6. Remoção de palavras indesejadas	21
3.7. Análise estatística	22
4. RESULTADOS.....	23
4.1. Primeira pergunta de pesquisa: variação entre vocábulos	23
4.1.1. Vocábulos mais utilizados no domínio empresarial	23
4.1.2. Vocábulos mais utilizados por usuários “influenciadores”	24
4.1.3. Vocábulos mais utilizados no domínio governamental.....	24
4.1.4. Vocábulos mais utilizados por usuários “aleatórios”	25
4.2. Segunda pergunta de pesquisa: variação entre domínios.....	25
4.3. Terceira pergunta de pesquisa: variação entre usuários.....	27
4.4. Quarta pergunta de pesquisa: Variação Temporal	28
5. CONSIDERAÇÕES FINAIS	30
REFERÊNCIAS	32
ANEXOS.....	34

1. INTRODUÇÃO

O inglês é a língua principal de comunicação entre povos diferentes do mundo, permitindo que pessoas falantes de línguas diferentes se comuniquem utilizando uma língua em comum. A língua inglesa atingiu esse estatuto de língua mundial devido à dominação política, econômica e cultural dos países anglófonos, especialmente o Reino Unido e Estados Unidos.

Como o inglês é a língua franca global (NORTHRUP, 2013), o uso de anglicismos, isto é, vocábulos e expressões originários da língua inglesa, tem se tornado comum em diversas línguas do mundo. Muitas expressões inglesas incorporadas a outras línguas advêm da ciência, negócios e tecnologia; como exemplos, podemos citar 'monkey pox' para 'varíola dos macacos', 'invoice' para 'fatura' e 'notebook' no sentido de 'computador portátil'.

Se a globalização e avanços tecnológicos vieram para fazer parte do nosso cotidiano, nos proporcionando não só maior praticidade e facilidade em diversos aspectos da nossa vida, mas também uma maior comunicação e acesso à informação a nível mundial, então também veio o anglicismo. A grande quantidade de anglicismos no dia a dia da população brasileira já por si só é um indício do quão difundida é a língua inglesa no vocabulário dos brasileiros e, aliado à democratização do acesso às redes sociais, ou seja, ao aumento dos usuários, a disseminação de opiniões cresceu paralelamente. Imagina-se, conseqüentemente, que o uso de termos e expressões em inglês também teve um crescimento, no mínimo similar, dentro do vocabulário informal das redes sociais.

Nos anos mais recentes, a população mundial infelizmente se viu obrigada a encarar a pandemia do COVID-19, que não só levou embora mais de 6 milhões de seres humanos, como também impôs diversas mudanças no estilo de vida de todos, desde o isolamento obrigatório até a adaptação das empresas para atender questões sanitárias. Seja no confinamento do *lockdown*, fazendo pagamentos com cartões *contactless*, ou assistindo séries no *streaming*, a questão da contribuição da pandemia para o uso do anglicismo no Brasil também deve ser levada em conta para esse tipo de pesquisa.

No entanto, o inglês não é usado somente para comunicação entre falantes de línguas diferentes, tendo também se incorporado em outras línguas. Assim, em muitas línguas diferentes empregam-se vocábulos oriundos da língua inglesa, muitas vezes no lugar de palavras já existentes no vernáculo. Por exemplo, no Brasil, é comum usar ‘off’ com o sentido de ‘desconto’, como em ‘10% off’ no lugar de ‘10% de desconto’, ‘backup’, como em ‘vou fazer um backup’ em vez de ‘vou fazer uma cópia de segurança’, entre tantos outros casos.

A língua inglesa atingiu esse estatuto de língua mundial por meio da dominação política, econômica e cultural dos países anglófonos, especialmente o Reino Unido e Estados Unidos. Ao mesmo tempo em que o inglês se torna uma língua tão influente, surgem movimentos de resistência à utilização do inglês em muitos países, como França e Brasil, entre outros. No caso do Brasil, questiona-se o emprego ‘exagerado’ ou ‘indiscriminado’ de ‘estrangeirismos’, o que estaria colonizando e desvirtuando a língua portuguesa, bem como influenciando negativamente na cultura nacional.

A pandemia do COVID-19 atingiu o planeta de forma devastadora, não só tirando a vida de milhões de seres humanos (MATHIEU *et al*, 2020), como também influenciando a rotina diária de todos os outros. Dentre outras coisas, os seres humanos foram forçados a lidar com maior frequência com meios tecnológicos para se manterem isolados ou para prevenir contato entre pessoas ou com superfícies contaminadas, como por exemplo cartões de crédito por aproximação, sensores automáticos para higienização das mãos e videoconferência, entre muitos outros. Uma consequência da pandemia que é pouco mencionada é a possível influência do isolamento social e do uso de tecnologias de comunicação remota, como as redes sociais, no uso de palavras estrangeiras. Mais especificamente, pode ter ocorrido um aumento do uso de palavras do inglês nas redes sociais durante a pandemia, porque as redes sociais são especialmente suscetíveis a culturas estrangeiras, por sua própria natureza. O inglês é a língua mais usada e mais influente nas redes sociais. As redes sociais são um meio de comunicação de larga escala, muitas vezes em locais distantes, ao redor do mundo. Devido ao caráter anglocêntrico das redes sociais bem como do isolamento social causado pela pandemia, é possível que tenha havido um incremento no uso de palavras inglesas na comunicação em português do Brasil nas redes sociais. A presente pesquisa pretende verificar essa hipótese por meio da coleta e análise de um grande corpus de postagens em português brasileiro no Twitter. Além de verificar se a

hipótese de trabalho é plausível, a mensuração do emprego de inglês na comunicação em rede social no país pode servir como um termômetro do alcance do uso de estrangeirismos em inglês na língua portuguesa do Brasil.

Assim, a presente pesquisa busca responder questionamentos como qual é a incidência de palavras em inglês e quais são essas palavras, quando usadas na língua portuguesa pelos brasileiros. Para tanto, um corpus de pesquisa foi coletado com as seguintes características: usuários brasileiros da rede social *Twitter*, falantes da língua portuguesa, ligados aos domínios de empresas, do governo, de influenciadores e de pessoas comuns.

O repertório em língua inglesa foi definido a partir de dois corpora já existentes, nomeadamente o Corpus iWeb, da língua inglesa; e o Corpus Literário de Machado de Assis (CLIMA) e o Corpus Literário Congênere (CLIC), de língua portuguesa.

Os corpora foram coletados e processados por meio de um script desenvolvido pelo professor orientador. Em seguida, foi feita análise estatística dos resultados empregando o pacote SAS OnDemand for Academics, por meio de um SAS Program criado pelo professor orientador.

A base teórico-metodológica do trabalho é a Linguística de Corpus, uma área dos estudos da linguagem que se ocupa da coleta e análise de grandes quantidades de texto armazenados na forma de arquivos de computador (BERBER SARDINHA, 2004). A seção seguinte do relatório apresenta e discute os pontos principais da Linguística de Corpus.

Esta pesquisa tem como objetivo: (a) descobrir a extensão da incidência de palavras da língua inglesa no uso cotidiano do português do Brasil na rede social *Twitter*; e (b) verificar a extensão da variação entre domínios/setores sociais e econômicos, entre usuários e temporal no uso dessas palavras, tendo como ponto de referência cronológica os anos em torno da pandemia de COVID-19.

Com base nos objetivos explicitados anteriormente, as seguintes perguntas foram elaboradas:

- i. Quais as médias de uso de palavras em inglês mais usadas no *Twitter* em português do Brasil?
- ii. Existe diferença na incidência de uso de palavras em inglês entre domínios/setores sociais e econômicos do *Twitter* no Brasil?
- iii. Existe diferença na incidência entre usuários em relação ao uso de palavras em inglês no *Twitter* em português do Brasil?
- iv. Existe variação temporal no uso de palavras em inglês no *Twitter* do Brasil? Mais especificamente, existe um aumento ou diminuição no uso de palavras em inglês tendo como ponto de referência a pandemia de COVID-19?

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção será apresentada a base teórica para a realização desta pesquisa.

2.1. Linguística de corpus (LC)

A Linguística de Corpus refere-se à área dos Estudos Linguísticos voltada para a coleta e análise de corpora, ou seja, coletâneas de texto falados, escritos, sonoros e visuais, por meio de tecnologia computacional (BERBER SARDINHA, 2004). Então ao falar de Linguística de Corpus, deve-se primeiramente definir o significado de “corpus”. Emprestada do latim, no qual significa “corpo”, corpus (e seu plural, corpora) é um conjunto de dados linguísticos escritos ou falados representando uma ou mais línguas, em um formato que possa ser lido por um computador (BERBER SARDINHA, 2004). Porém, diferentemente de outros conjuntos de dados, o corpus é construído com objetivos específicos a partir de textos naturais, ou seja, textos existentes na língua, que não foram produzidos artificialmente ou com o específico propósito de fazer parte do corpus. Resumidamente, a definição abaixo pode ser considerada a definição mais completa, uma vez que está de acordo com as diversas características de um corpus descritas (BERBER SARDINHA, 2004, p. 18-19):

A origem: os dados devem ser autênticos. O propósito: o corpus deve ter a finalidade de ser um objeto de estudo linguístico. A composição: o conteúdo do corpus deve ser criteriosamente escolhido. A representatividade: o corpus deve ser representativo de uma língua ou variedade. A extensão: o corpus deve ser vasto para ser representativo.

Primeiramente, é possível categorizar o corpus por meio de sua tipologia que, apesar de extensa, pode ser descrita de acordo com os seguintes critérios: um corpus pode ser visto por variáveis como: modo, isto é, se é composto de textos que refletem o modo verbal, falado, sonoro ou visual; tempo, se representa períodos de tempo distintos; por seleção (corpora de amostragem, monitor, dinâmico, estático ou equilibrado); por conteúdo (quando os textos são de gêneros específicos, de uma ou mais variedades sociolinguísticas específicas ou de idiomas diferentes); por autoria (se os autores do texto são falantes nativos ou não); pelo tipo de disposição interna do corpus; e por sua finalidade (corpora de estudo, de referência ou de teste).

Um objetivo central de um corpus é ser “representativo da linguagem, de um idioma, ou de uma variedade dele.” (BERBER SARDINHA, 2004, p. 22). E apesar de não existir critérios muito objetivos para determinar a representatividade de um corpus, existe, todavia, a necessidade de possuir a maior quantia possível de dados linguísticos, a fim de obter uma amostra abundante de características linguísticas, especialmente de baixa frequência. Por exemplo, pesquisadores que desejam analisar a variação lexical presente nos sambas-enredo de escolas dos estados de São Paulo e Rio de Janeiro necessitam de um corpus que contenha em sua totalidade, e quando não for possível, predominantemente, dados linguísticos em língua portuguesa escritos ou falados por brasileiros nativos dessas regiões, dentro do contexto proposto, isto é, sambas-enredo (vide GONÇALVES, 2019). No caso desta pesquisa, o corpus de estudo teve o propósito de representar a língua portuguesa empregada no Twitter por setores específicos (domínios). Assim, o corpus coletado e analisado nesta pesquisa é composto por textos escritos, somente de falantes nativos da língua portuguesa, durante um único período; equilibrado (como explicitado no item 3.1 adiante); de estudo e; de gêneros específicos, já que é constituído apenas de postagens do Twitter do Brasil em língua portuguesa.

De acordo com Berber Sardinha (2004, p. 35), existe um debate sobre se a LC seria uma disciplina ou metodologia. Talvez ela possa ser mais facilmente definida ao pensarmos no que ela não é capaz de fazer. De acordo com Bennett (2010), a LC não é capaz, por exemplo, de prover evidências negativas, isto é, o corpus não irá nos dizer se algo está certo ou se é possível, ou não, em algum aspecto da língua; o corpus pode apenas afirmar o que está presente ou não presente em sua constituição. Além disso, ainda segundo Bennet (2010), o corpus não irá justificar a disposição de seus dados, cabe ao pesquisador, baseado nos propósitos da pesquisa, fazer a análise daquilo que o corpus apresenta. Outro detalhe que a autora chama atenção é que o corpus nunca será representativo de toda a linguagem, por maior que seja, nem todas as instâncias de uso da língua estarão presentes.

De acordo com Kennedy (2014, tradução do autor), “seria errado sugerir que a linguística de corpus é uma teoria da linguagem paralela a outras teorias como a da gramática transformacional, ou até mesmo que ela é um braço novo ou separado da linguística.”¹ Por um lado, a Linguística de Corpus seria uma “metodologia”, visto que ela comporta

¹ Traduzido do original “It would be misleading, however, to suggest that corpus linguistics is a theory of language in competition with other theories of language such as transformational grammar, or even more that it is a new or separate branch of Linguistics.” (KENNEDY, 2014)

metodologias próprias que podem ser usadas com diversas teorias. Por outro lado, precisamos considerar que a LC possa não ser nem mesmo uma metodologia ou teoria. Berber Sardinha (2014, p. 37) considera que a Linguística de Corpus seria “uma maneira de se chegar à linguagem”, ou como alguns linguistas preferem, uma “abordagem baseada em corpus”. Com base em dados linguísticos compilados em um corpus, um pesquisador pode analisar a variação do discurso da língua, como no caso do corpus presente nesta pesquisa, que possui dados para uma análise da variação temporal e entre grupos do uso de anglicismos dentro de um contexto específico. Pensando no corpus criado para a pesquisa do presente trabalho, podemos sugerir que ele forma a base de uma abordagem linguística empírica de descoberta do alcance do uso de anglicismos na língua portuguesa usada do Twitter do Brasil.

Mesmo não sendo possível determinar categoricamente se a LC seria uma teoria ou metodologia, visto que essa questão extrapola o escopo desta pesquisa, é importante ressaltar o caráter empírico da pesquisa: a pesquisa levada a cabo tem objetivos e perguntas de pesquisa voltados para a análise de dados. A presente pesquisa não tem objetivos teóricos, voltados para a discussão teórica das questões envolvidas na presença de estrangeirismos na língua portuguesa. A discussão teórica, que é claramente importante, será enfocada em trabalhos decorrentes desta pesquisa, como artigos, apresentação em congresso e estudos posteriores em nível de pós-graduação.

Visto que a Linguística de Corpus está intrinsecamente atrelada ao avanço tecnológico, sua história mudou consideravelmente com a maior capacidade de processamento e disponibilidade de equipamentos. Antes da invenção do computador, a exploração e o armazenamento de corpora se davam de forma manual.

“A não ser que usemos um computador para ler, procurar e manipular os dados, trabalhar com conjuntos de dados extremamente grandes não é viável, por causa do tempo que um analista humano ou um grupo de analistas demoraria para procurar nos textos.” (MCENERY, 2012, p. 2, tradução do autor)²

Como visto anteriormente, o pesquisador busca ter em seu corpus a maior quantidade possível de dados, portanto a presença de meios computacionais para coleta faz com que a

² Traduzido do original: “Unless we use a computer to read, search and manipulate the data, working with extremely large datasets is not feasible because of the time it would take a human analyst, or team of analysts, to search through the text.” (MCENERY, 2012, p.2)

quantidade de dados aumentasse consideravelmente. Alguns marcos tecnológicos da Linguística de Corpus são: o SEU (Survey of English Usage), um corpus não-computadorizado, de textos escritos e falados, com cerca de 1 milhão de palavras do inglês britânico, famoso por sua composição organizada, servindo de referência para o desenvolvimento dos etiquetadores, no fim dos anos de 1950; o “primeiro etiquetador morfossintático para computador” de Taggit em 1970; “um dos programas pioneiros para microcomputadores” em 1987, que permitiu a listagem de palavras e concordâncias; e o conjunto de ferramentas WordSmith Tools, primeiro a levar a Linguística de Corpus para o ambiente Windows, já no fim do século XX (cf. BERBER SARDINHA, 2004).

Em relação a corpora importantes na história da Linguística de Corpus, podemos citar o Brown University Standard Corpus of Present-day American English, corpus pioneiro com 1 milhão de palavras escritas no inglês americano, de 1964; o British National Corpus, o primeiro a conter 100 milhões de palavras do inglês britânico, de 1995; e o Bank of English, um corpus de palavras do inglês britânico, de 1987 (cf. BERBER SARDINHA, 2004).

Já no Brasil, Biderman (1978) cita alguns corpora da língua portuguesa como pioneiros: o Frequency dictionary of Portuguese words, um dos primeiros corpora eletrônicos de português e os de português do Brasil compilados por Jean Roche (década de 1960) e J. Hutchins (década de 1970), além do corpus compilado por ela mesma, em 1969.

2.2. Estrangeirismos

Estrangeirismo é “o emprego, na língua de uma comunidade, de elementos oriundos de outras línguas” (FARACO, 2001, p. 15). O uso de palavras e expressões estrangeiras no português brasileiro, como em qualquer língua, não é um fenômeno recente. Entre os séculos XIX e XX, na língua portuguesa empregavam-se muitas palavras do francês (chamados de galicismos ou francesismos), a língua franca da época. Desde meados do século XX, com a maior influência da língua inglesa, devido à hegemonia econômica norte-americana, o uso de estrangeirismos do inglês, chamados de anglicismos, passou a ser abundante na língua portuguesa, bem como em outras línguas. Segundo Santos (2006), atualmente a maioria dos estrangeirismos em inglês seria de termos ligados à tecnologia, visto que a tecnologia é desenvolvida no exterior e recebe termos em inglês, na origem.

Podemos encontrar estrangeirismos no português brasileiro em palavras aportuguesadas, como “abajur” (do francês *abat-jour*), “bife” (do inglês *beef*), “estrogonofe” (do russo *stroganov*) e “muçarela” (do italiano *mozzarella*); assim como palavras não aportuguesadas,

como *delivery* (palavra inglesa para “entrega”), *on-line* (palavra inglesa para “conectado”) e *réveillon* (palavra francesa para “Ano-Novo”).

Devido à percepção de alguns setores da sociedade de que os estrangeirismos estivessem sendo usados em demasia, surgiram movimentos voltados à contenção do uso de palavras e expressões estrangeiras como uma medida de proteção e preservação da identidade cultural nacional, como os projetos de lei de Aldo Rebelo (BRASIL, 1999) e Raul Carrion (BRASIL, 2009). Tais iniciativas alegam que existe uma “invasão indiscriminada e desnecessária de estrangeirismos” (REBELO, 1999, p.107), de palavras usadas em inglês, mesmo quando existem vocábulos equivalentes no português, como ‘*drink*’ no lugar de ‘aperitivo’ e de palavras aportuguesadas, como por exemplo “printar” (verbalização do substantivo em inglês *print*).

No entanto, todas as línguas vivas acabam tendo contato com outras, resultando na incorporação de empréstimos, o que Faraco (2001) entende como um processo natural da globalização. Além disso, o estrangeirismo, segundo Bosi (2017, p.11), não apenas é inevitável, como “não afeta a língua portuguesa”, já que ocorreria “um acréscimo no vocabulário”, ou seja, o uso de estrangeirismos enriqueceria a língua portuguesa.

3. METODOLOGIA

Nesta seção serão apresentados os procedimentos metodológicos utilizados durante a pesquisa.

3.1. Design e coleta do corpus de pesquisa

O corpus de estudo desta pesquisa consistiu em uma coletânea do componente verbal (mensagens digitadas) de postagens do *Twitter* em língua portuguesa.

O primeiro passo para a construção do corpus de pesquisa foi decidir quais contas teriam as postagens incluídas. Assim, a amostragem foi organizada em quatro grupos, ou em tipos de usuários, por assim dizer. Ao todo foram 200 usuários de *Twitter*, os quais foram categorizados em: 50 usuários “governamentais” ou relacionados ao governo, seja federal, estadual ou municipal; 50 usuários “empresariais” de diferentes atuações no mercado, separados por seu ramo de acordo com o Anexo A; 50 usuários “influenciadores”, ou seja, usuários de pessoas famosas com bastante engajamento com seus seguidores; e 50 usuários “aleatórios”, ou seja, usuários de pessoas comuns, consideradas a população brasileira no geral.

Todos as postagens ocorreram no intervalo de tempo entre janeiro de 2019 e outubro de 2022, pois este período engloba o ano anterior à pandemia e os anos de sua duração, até o começo da coleta dos dados da pesquisa.

3.1.1. Critérios para seleção dos usuários

A coleta dos usuários seguiu alguns outros critérios que serão listados a seguir:

- i. Como o ano de 2022 é ano de eleição no Brasil, muitos usuários “governamentais” que poderiam fazer parte do corpus foram suspensos ou foram substituídos por usuários provisórios, portanto esses usuários foram descartados por não se encaixarem no escopo da coleta do corpus de pesquisa. O motivo de tal ação federal foi uma forma do governo se adequar ao artigo 73 da Lei 9.504/1997 que dispõe das “Condutas Vedadas aos Agentes Públicos em Campanhas Eleitorais” onde, no item 6, que aborda as regras que devem ser cumpridas a 3 meses do 1º turno, a legislação diz:

[...] com exceção da propaganda de produtos e serviços que tenham concorrência no mercado, autorizar publicidade institucional dos atos, programas, obras, serviços e campanhas dos órgãos públicos federais, estaduais ou municipais, ou das respectivas entidades da administração indireta, salvo em caso de grave e urgente necessidade pública, assim reconhecida pela Justiça Eleitoral;

- ii. A obtenção dos usuários “influenciadores” teve por base a lista de usuários brasileiros com mais seguidores no *Twitter* no ano de 2022.
- iii. A escolha dos usuários “empresariais” se deu por usuários que tivessem um engajamento alto com seus seguidores.
- iv. Quanto aos usuários “aleatórios”, a escolha se deu com base em postagens consideradas “viralizadas”, ou seja, postagens que atingiram um número alto de curtidas, retuítes e comentários, i.e. um engajamento alto.
- v. Todos os usuários passaram por um filtro manual onde se checava se o usuário:
 - a) tinha um número considerável de postagens dentro do período escolhido para o corpus; b) não possuía postagens somente em língua inglesa (a fim de atingir pessoas do mundo todo); e c) não eram páginas de fã clube ou páginas controladas por inteligência artificial (chamadas de *bots*).

3.2. *Scraping*

Uma vez consolidada a lista de usuários dos quais seriam coletados os tweets, foi necessário fazer o *scraping* das mensagens, ou seja, a raspagem dos dados da internet. a coleta automatizada em lote. Tal ação é uma forma de extrair dados de sites da web convertendo-os em informação para uma análise posterior. Para isso, o professor orientador desenvolveu, especificamente para esta pesquisa, um *script*, ou seja, um roteiro em linguagem de programação que automatiza a execução das tarefas necessárias para a obtenção dos dados. Esse processamento se deu no sistema operacional Ubuntu, um sistema Linux. O *scraping* foi realizado por meio da ferramenta *snsrape*, que especificamente coleta postagens em massa de redes sociais como Facebook, Instagram, Reddit e *Twitter*, retornando as postagens em arquivos no formato JSON. Esses arquivos trazem dados como o usuário, a mensagem, *hashtags*, data, horário, etc.

No caso desta pesquisa, os dados foram salvos pelo script em arquivos separados por categoria, usuário, ano e trimestre (janeiro a março, abril a junho, julho a setembro e outubro a dezembro).

3.3. Limpeza de corpus

O passo seguinte é o de limpeza do corpus, ou seja, na conversão dos arquivos brutos JSON em arquivos limpos em TXT. Os arquivos JSON são um formato de saída em XML, com visualização complicada para o olhar humano. A limpeza do corpus elimina esse código XML e retém apenas os dados desejados: o conteúdo das postagens; a data da postagem; e o usuário de quem fez a postagem. Esse processo de limpeza foi realizado por um script desenvolvido pelo professor orientador.

3.4. Corpus de Referência

A fim de detectar a presença de palavras inglesas nas postagens em português, foi preciso empregar dois corpora de referência: um corpus em inglês e outro corpus em português.

O primeiro corpus, conforme mostra a Tabela 1, é o corpus iWeb (DAVIES; KIM, 2019, de textos em língua inglesa. O segundo, conforme a Tabela 2, é o Corpus Literário de Machado de Assis (CLIMA), que abrange romances e contos do escritor, e o Corpus Literário Congênere (CLIC), composto por obras de escritores do período de 1850 a 1910 (KAUFFMANN, 2020; KAUFFMANN; BERBER SARDINHA, 2021).

Tabela 1 — Design do corpus iWeb

Textos (páginas da web)	22.388.141
Tokens	≈ 14.000.000.000
Websites	94.391

Fonte: DAVIES; KIM, 2019

Tabela 2 — Design do corpus CLIMA-CLIC

Textos	228
Tokens	4.359.478
Types	11.776

Fonte: KAUFFMANN, 2020; KAUFFMANN; BERBER SARDINHA, 2021.

3.5. Indexação do corpus

Na criação do corpus, foi necessário separar o conteúdo específico para análise de toda a informação limpa obtida no processo de *scraping*. Para tanto, os dois corpora mencionados no item anterior foram utilizados: (I) o corpus iWeb, de palavras em língua inglesa, como base para a localização das palavras que irão formar o corpus; (II) o corpus CLIMA-CLIC, de palavras em língua portuguesa, que irá servir como filtro para remover as palavras que não irão formar o corpus, ou seja, as palavras em língua portuguesa.

A partir disso, o script criou um outro arquivo TXT com a lista de palavras em inglês presentes em todas as postagens de todos os usuários selecionados, com base no corpus I e II.

O algoritmo usado no script para detecção de palavras estrangeiras nos tweets é o seguinte:

- i. Extrair uma lista de palavras com as 200 mil palavras mais frequentes do iWeb. Essa lista representa o vocabulário em inglês contemporâneo em uso ao redor do mundo, por meio de script especializado.
- ii. Extrair uma lista de todas as palavras do corpus CLIMA-CLIC. Essa lista representa a língua portuguesa sem a presença de estrangeirismos recentes como *site*, *backup* e *download*, por meio de script especializado.
- iii. Comparar a lista do iWeb com a lista do CLIMA-CLIC. O resultado é uma lista de candidatos a estrangeirismos, por meio de script especializado.
- iv. Verificar, manualmente, a lista de candidatos, extraindo formas vocabulares cognatas, como *dental*, *legalize*, *maestro*, *minimize*, *informal* e *impasse*; falsos cognatos tais como *doer*, *ranger* e *clique*; e formas do inglês que, por terem sido mal acentuadas, passam a ter correspondência em português, como *'tripe'* (*tripé*), *'bone'* (*boné*) e *nebula* (*nébula*).
- v. Identificar a presença de cada palavra da lista de candidatos final em cada postagem, por meio de script especializado.

3.6. Remoção de palavras indesejadas

Uma vez estando pronta a lista de palavras em língua inglesa, iniciou-se a remoção de palavras indesejadas. Essas palavras são consideradas “falsos cognatos” ou “falsos positivos”, isto é, são palavras que possuem a mesma grafia tanto em inglês quanto em

português. Além disso, palavras dessa lista consideradas “nomes próprios” também foram removidas.

Com isso, o corpus é alterado de duas formas: negativamente, pois se retira os casos em que as palavras foram utilizadas em seu significado da língua inglesa e positivamente, pois assim também se retira as palavras presentes no corpus I que foram utilizadas em seu significado da língua portuguesa.

Durante a remoção das palavras indesejadas, o autor entendeu que tal processo tem um impacto com um peso positivo maior, em razão da maioria das palavras removidas serem mais utilizadas no português do que no inglês.

3.7. Análise estatística

A fim de comparar os dados disponíveis no corpus coletado em relação aos domínios e aos períodos de tempo, foi empregado o pacote estatístico SAS OnDemand for Academics (https://www.sas.com/en_us/software/on-demand-for-academics.html). Mais especificamente, foi criado um script SAS (SAS Program; vide Anexo B) para realizar o teste F, que faz a comparação estatística entre as médias dos grupos envolvidos, de forma com que as perguntas pudessem ser respondidas objetivamente.

Em síntese, o teste F compara duas médias para saber se existe uma diferença real entre elas, no qual o valor de F representa a magnitude da diferença e o valor de p representa o valor da significância estatística. Para começar a assumir como verdadeiro o fato de existir uma diferença entre as médias, o valor de p precisa ser menor que 0,05 e, quanto menor seu valor e maior o valor de F, maior a diferença entre as médias.

4. RESULTADOS

Nesta seção serão apresentados os resultados e as respostas para as perguntas propostas nesta pesquisa.

4.1. Primeira pergunta de pesquisa: variação entre vocábulos

Para poder responder a essa pergunta, primeiramente foi feito o levantamento das palavras usadas por cada domínio. Os resultados são apresentados a seguir:

4.1.1. Vocábulos mais utilizados no domínio empresarial

No domínio empresarial, foram empregados 2369 vocábulos em inglês. Conforme aparece na Tabela 3 a seguir, a presença de termos associados a informática e a vendas ou propagandas é bastante notória, em consequência de o *Twitter* ser também um canal de contato direto das empresas com os consumidores, como em *app*, *email*, *promo* e *inbox*.

Tabela 3 — Os 10 vocábulos mais usados em inglês no domínio empresarial

Palavra	Frequência
app	170738
site	83650
email	73608
hey	32908
off	27491
big	20971
promo	11598
online	11104
inbox	10959
tweet	10894

Fonte: o Autor.

4.1.2. Vocábulo mais utilizados por usuários “influenciadores”

No domínio dos famosos, foram empregados 1404 vocábulos em inglês. Conforme aparece na Tabela 4 a seguir, podemos perceber o uso de palavras cotidianas da esfera das redes sociais, como também palavras usadas no mundo das celebridades, como por exemplo *tweet*, *timeline*, *reality* e *hit*.

Tabela 4 — Os 10 vocábulos mais usados em inglês no domínio dos famosos

Palavra	Frequência
tweet	980
site	569
look	388
app	354
timeline	341
lady	281
reality	273
news	264
feat	238
hit	215

Fonte: o Autor.

4.1.3. Vocábulos mais utilizados no domínio governamental

No domínio do governo, foram empregados 889 vocábulos em inglês. Conforme aparece na Tabela 5 a seguir, o predomínio é de palavras relacionadas à internet e ao acesso, provavelmente, de serviços do governo, como *webinar*, *site*, *app* e *links*.

Tabela 5 — Os 10 vocábulos mais usados em inglês no domínio do governo

Palavra	Frequência
site	4207
online	2403
email	1287
app	984
news	799
info	310
webinar	243

sites	215
links	206
web	198

Fonte: o Autor.

4.1.4. Vocábulo mais utilizados por usuários “aleatórios”

Por fim, no domínio de pessoas comuns, foi empregado o maior vocabulário, com 3360 vocábulos em inglês. Conforme aparece na Tabela 6 a seguir, a predominância é novamente de palavras usadas nas redes sociais, assim como no domínio das celebridades, como *tweet*, *fav*, *app* e *off*.

Tabela 6 — Os 10 vocábulos mais usados em inglês no domínio de pessoas comuns

Palavra	Frequência
tweet	4757
site	2587
old	1827
online	1381
fav	1172
app	1024
off	1020
tweets	1003
love	917

Fonte: o Autor.

4.2. Segunda pergunta de pesquisa: variação entre domínios

A Tabela 7 a seguir traz as médias por domínio. O teste F apresenta que houve uma diferença significativa entre os domínios ($F=15466,9$, $p<0,0001$). O domínio com maior média de uso é o empresarial, o que confirma nossa intuição, uma vez que é esperado das empresas uma alta frequência de postagens e engajamento com seus clientes. No entanto, o fato de o grupo de pessoas comuns ser o segundo colocado é fora do esperado, mesmo utilizando em média metade da quantidade de palavras em inglês do grupo das empresas. O fato de o grupo dos influenciadores ter ficado abaixo do grupo de pessoas comuns também é surpreendente, visto que esperávamos que eles usassem o inglês mais frequentemente.

Porém, o grupo do governo ser o que menos emprega palavras em inglês apareceu como esperado. Uma vez que a linguagem do governo precisa ser de entendimento geral da população, presumimos que esse domínio teria a maior predominância da língua portuguesa, utilizando-se resumidamente de palavras relativas ao uso da internet.

Tabela 7 — Média de uso de palavras inglesas por domínio.

Domínio	Palavras em inglês	Média	Desvio padrão
Empresas	3856915	0,17168359	0,44755107
Governo	349372	0,06411504	0,27165230
Influenciadores	233767	0,07847558	0,33772780
Pessoas comuns	1175107	0,09887866	0,42385360

Fonte: o Autor.

Além das médias, fizemos um levantamento por outra perspectiva, levando em conta a quantidade de postagens que contém pelo menos uma palavra em inglês, conforme os resultados mostrados na Tabela 8 a seguir:

Tabela 8 — Postagens com palavras inglesas por domínio.

Domínio	Palavras em inglês	Tweets	Razão	Tweets com palavras em inglês
Empresas	564235	3856915	6,8	14,6%
Governo	20335	349372	17,2	5,8%
Influenciadores	15072	1175107	78,0	1,3%
Pessoas comuns	86516	233767	2,7	37,0%

Fonte: o Autor.

Esses resultados mostram que o grupo com maior número de *tweets* com palavras em inglês é o de pessoas comuns (37%), seguido dos grupos das empresas, do governo e dos influenciadores. Em termos de incidência de uso de palavras em inglês, há uma postagem com alguma palavra em inglês em cada 2,7 postagens no grupo de pessoas comuns, seguido de 1 em cada 6,8 para empresas, 1 em cada 17,2 para o governo e 1 em cada 7,68 para influenciadores.

Comparando as Tabelas 7 e 8, podemos ver que as empresas, embora utilizem uma quantidade maior de palavras em inglês, publicam uma quantidade menor de postagens com alguma palavra em inglês, do que as pessoas comuns. Os usuários “aleatórios”, pelo contrário, usam com mais frequência palavras em inglês, apesar de usar uma quantidade menor de palavras. Isso é de se esperar, dado ao contexto limitado que algumas postagens

das empresas podem conter, como uma resposta para o cliente, onde se deve manter um nível de formalidade e passar as informações necessárias de forma mais clara possível.

4.3. Terceira pergunta de pesquisa: variação entre usuários

Em relação à segunda pergunta de pesquisa, a Tabela 9 traz os vinte usuários com maiores médias de uso de palavras em inglês no corpus. O teste F mostra que houve diferença significativa entre os usuários ($F=1130,30$, $p<0,0001$). No Anexo C é apresentada a relação completa com os 200 usuários. Ao final do nome do usuário, incluímos um código referente ao domínio a que o usuário está ligado: *_em* para empresas, *_go* para governo, *_in* para influenciadores e *_pc* para pessoas comuns/aleatórias.

Tabela 9 — Relação dos 20 usuários com maior média de uso de palavras em inglês

Usuário (<i>_domínio</i>)	Palavras em inglês	Média	Desvio padrão
pierre_00_pc	9277	0,54500377	1,13199866
riachuelo_em	46067	0,51509758	0,81771312
playstation_br_em	8616	0,50394615	0,86122817
cea_brasil_em	37085	0,40461103	0,64412020
mmamelotto_pc	14449	0,40141186	0,95546641
azulinhasaereas_em	13516	0,39124001	0,62841360
amazonbr_em	844	0,37796209	0,67353822
positivooficial_em	5067	0,37418591	0,66562775
xboxbr_em	7045	0,36579134	0,79699497
avonbr_em	9216	0,34114583	0,61341819
citroenbrasil_em	921	0,33876221	0,71193718
oficialpeugeot_em	2656	0,32228916	0,63486893
googlebrasil_em	2665	0,32007505	0,61205850
microsoftbr_em	4404	0,31607629	0,67397381
renaultbrasil_em	4642	0,30978027	0,69791081
deezerbr_em	10057	0,30615492	0,54967527
kiadobrasil_em	536	0,29850746	0,56077184
lojas_renner_em	33222	0,29787490	0,59684787
santander_br_em	186467	0,29156902	0,51690490
nissanbrasil_em	2289	0,28877239	0,49389695

Fonte: o Autor.

Os resultados mostram que há somente dois representantes do grupo das pessoas comuns na relação dos vinte usuários que mais utilizam palavras em inglês, conforme destacado na tabela anterior.

4.4. Quarta pergunta de pesquisa: Variação Temporal

Em relação à quarta pergunta de pesquisa, a Tabela 10 a seguir traz as médias por ano. O teste F mostra que houve diferença significativa entre os quatro anos estudados ($F=1816,95$, $p<0,0001$). A média de uso de palavras inglesas cresceu de 0,13 em 2019 para 0,15 em 2021 e para 0,16 em 2021.

Em 2022, houve decréscimo da média para 0,13, patamar semelhante a 2019. Os desvios padrão são altos em relação à média, indicando que há muita variação dentro de cada ano. Em geral, os resultados indicam que houve um aumento no uso de palavras em inglês durante a pandemia de COVID-19.

Tabela 10 — Média de uso de palavras inglesas em cada ano

Ano	Palavras em inglês	Média	Desvio padrão
2019	983502	0,13111005	0,40877088
2020	1588476	0,15101771	0,43088603
2021	1877968	0,16027270	0,45353762
2022	1165215	0,12811799	0,41232776

Fonte: o Autor.

Ainda em relação a essa mesma pergunta, foram calculadas as médias de uso por trimestre de cada ano, conforme mostra a Tabela 11 abaixo. Segundo o teste F, houve diferença estatística significativa entre os trimestres ($F=767,22$, $p<0,0001$), sendo os trimestres com maior média de uso de palavras em inglês o terceiro semestre de 2021 (média de 0,18), seguido do segundo de 2020 (0,17) e o terceiro de 2020 (0,16). Isso confirma o resultado anterior, mostrando que os trimestres com maior média de uso de palavras em inglês ocorreram durante a pandemia de COVID-19.

Tabela 11 — Média de uso de palavras em inglês por trimestre

Trimestre	Palavras em inglês	Média	Desvio padrão
-----------	--------------------	-------	---------------

2019 01 03	233006	0,13402230	0,41101730
2019 04 06	233975	0,12399615	0,40517429
2019 07 09	237794	0,12646661	0,40781338
2019 10 12	278727	0,13860875	0,41055594
2020 01 03	320149	0,11189477	0,39104812
2020 04 06	416304	0,16831931	0,44332653
2020 07 09	405514	0,16158258	0,44783038
2020 10 12	446509	0,15334293	0,42894775
2021 01 03	484245	0,14195913	0,40957893
2021 04 06	469045	0,16486478	0,45325436
2021 07 09	475752	0,18052052	0,48914033
2021 10 12	448926	0,15377145	0,45876428
2022 01 03	419898	0,13147241	0,41550557
2022 04 06	356474	0,12549583	0,41162761
2022 07 09	335808	0,12326389	0,40917691
2022 10 12	53035	0,14991986	0,41073453

Fonte: o Autor.

5. CONSIDERAÇÕES FINAIS

Com o intuito de descobrir a extensão da presença e a variação no uso de anglicismos na língua portuguesa dentro da rede social Twitter por brasileiros, conduzimos uma pesquisa empírica com base em um grande corpus de tweets coletado criteriosamente para esta pesquisa e processado por meio de scripts desenvolvidos especialmente para a investigação.

Podemos dizer que, de modo geral, os resultados se aproximaram do esperado.

De acordo com os resultados, o domínio empresarial apresenta o maior número de palavras em inglês e a maior média de uso, majoritariamente associadas a informática e a vendas, devido à uma grande frequência de postagens. Apesar disso, o número de tweets com alguma palavra em inglês não é o maior, pois apresenta uma média de tweets de 14,6% contra 37,0% do domínio de pessoas comuns, que predominantemente emprega palavras usadas nas redes sociais e no mundo das celebridades.

O que não foi esperado, entretanto, foi a presença de dois usuários “aleatórios” na lista dos dez usuários com maior média de uso de anglicismos, da qual os oito restantes são usuários “empresariais”. Além disso, esperava-se que os usuários “influenciadores” apresentassem uma quantidade maior de uso do anglicismo em suas postagens, o que não foi confirmado.

Em relação ao uso do anglicismo de acordo com o tempo, a hipótese de que durante o foco da pandemia e do isolamento social o uso de palavras inglesas no *Twitter* do Brasil aumentou foi confirmada, uma vez houve um crescimento no uso no começo da pandemia (último trimestre de 2019 e primeiro de 2020) e uma redução no uso ao final (último trimestre de 2021 e primeiro de 2022), com seu pico em 2021.

Este trabalho pretende ter contribuído para preencher a lacuna relativa à falta de dados sobre a extensão do uso de anglicismos na língua portuguesa. Os dados apresentados mostram a incidência real dos anglicismos em uma variante da língua portuguesa, postagens do Twitter. Esses dados podem ser usados para informar o debate sobre estrangeirismos, dimensionando a extensão de sua presença bem como especificando os setores da sociedade onde mais e menos ocorrem.

Em trabalhos futuros, seria interessante comparar as médias de ocorrência obtidas em nosso estudo com médias de uso no português brasileiro em geral, a fim de sabermos se o uso de anglicismos no Twitter é maior ou menor do que na língua portuguesa do Brasil em geral.

Também seria interessante verificar os usos das palavras empregadas, para sabermos se os usos dessas palavras no português brasileiro e no inglês são semelhantes ou diferentes.

Essas e outras questões não puderam ser averiguadas neste estudo tendo em vista o escopo de uma pesquisa de TCC.

Este trabalho mostra que é possível (e desejado) usar a Linguística de Corpus para averiguar questões que pautam o debate público sobre a vida nacional, a fim de que os setores envolvidos tenham dados concretos de onde partir. No caso dos estrangeirismos, o levantamento de dados de larga escala como os nossos são fundamentais.

REFERÊNCIAS

BENNETT, Gena R. **Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers**. Michigan: ELT, 2010.

BERBER SARDINHA, Tony. **Linguística de Corpus**. 1ª ed. São Paulo: Manole, 2004.

BIDERMAN, M. T. C. **Teoria linguística: linguística quantitativa e computacional**. Rio de Janeiro/São Paulo: LTC, 1978.

BOSI, Ana Clara C. M. **Estrangeirismos: “up” na língua portuguesa ou “déficit” da identidade linguística?** Minas Gerais, 2017. Disponível em: <<http://repositorio.fucamp.com.br/bitstream/FUCAMP/213/1/Estrangeirismoslinguaportuguesa.pdf>>. Acesso em: 02 nov. 2022.

BRASIL, Assembleia Legislativa do Rio Grande do Sul. **Projeto de Lei Raul Carrion 156/2009**. Institui a obrigatoriedade da tradução de expressões ou palavras estrangeiras para a língua portuguesa, sempre que houver em nosso idioma palavra ou expressão equivalente, no âmbito do Estado do Rio Grande do Sul e dá outras providências. Porto Alegre, 5 de agosto de 2009. Disponível em: <<http://proweb.procergs.com.br/Diario/DA20090817-01-100000/EX20090817-01-100000-PL-156-2009.pdf>>.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 1.676/1999**. Dispõe sobre a promoção, a proteção, a defesa e o uso da Língua Portuguesa e dá outras providências. Brasília: Câmara dos Deputados, 1999. Disponível em: <<http://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=17069>>. Acesso em: 30 out. 2022.

BRASIL. Lei nº 9.504 de 30 de setembro de 1997. Estabelece normas para as eleições. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/19504.htm>. Acesso em: 04 nov. 2022.

DAVIES, Mark; KIM, Jong-Bok. **The advantages and challenges of “big data”**: Insights from the 14 billion word iWeb corpus. *Linguistic Research*, 36, n. 1, p. 1-34, 2019.

FARACO, Carlos Alberto. **Estrangeirismos: guerras em torno da língua.** São Paulo: Parábola, 2001.

GONÇALVES, Leticia. **Uma análise de sambas-enredo: A variação lexical a partir da Linguística de Corpus.** São Paulo, 2019.

KAUFFMANN, Carlos; BERBER SARDINHA, Tony. Brazilian Portuguese literary style. In: FRIGINAL, E. e HARDY, JACK. **The Routledge Handbook of Corpus Approaches to Discourse Analysis.** Abingdon: Routledge. p. 354-375. 2021.

KAUFFMANN, Carlos. **Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis [Corpus Linguistics and style: Multi-dimensional and canonical analyses of Machado de Assis's fiction].** (Tese de doutoramento) - LAEL, Programa de Pós-Graduação em Linguística Aplicada e Estudos da Linguagem, São Paulo. 2020.

KENNEDY, Graeme. **An introduction to Corpus Linguistics.** Inglaterra: Routledge, 2014.

MATHIEU, Edouard; RITCHIE, Hannah; RODÉS-GUIRAO, Lucas; APPEL, Cameron; GIATTINO, Charlie; HASELL, Joe; MACDONALD, Bobbie; DATTANI, Saloni; BELTEKIAN, Diana; ORTIZ-OSPINA, Esteban; ROSER, Max. **Coronavirus Pandemic (COVID-19) Deaths.** Our World Data, 2020. Disponível em: <<https://ourworldindata.org/coronavirus>>. Acesso em: 17 nov. 2022.

MCENERY, Tony; HARDIE, Andrew. **Corpus Linguistics: Method, Theory and Practice.** 1ª ed. United States of America: Cambridge University Press. 2012.

NORTHRUP, David. **How English Became the Global Language.** [S.l.]: Palgrave Macmillan. 2013.

SANTOS, Agenor Soares. **Dicionário de anglicismos e de palavras inglesas correntes em português.** Rio de Janeiro: Elsevier, 2006.

ANEXOS

ANEXO A – Relação de usuários do domínio empresarial

Empresa	Usuário	Ramo
Coca-Cola	cocacola_br	ALIMENTOS
Guaraná Antarctica	guarana	ALIMENTOS
McDonalds	mcdonalds_br	ALIMENTOS
Chevrolet	chevroletbrasil	AUTOMÓVEIS
Citröen	citroenbrasil	AUTOMÓVEIS
Nissan	nissanbrasil	AUTOMÓVEIS
Renault	renaultbrasil	AUTOMÓVEIS
Volkswagen	vwbrasil	AUTOMÓVEIS
Banco do Brasil	bancodobrasil	BANCO
Bradesco	bradesco	BANCO
Itaú	itau	BANCO
Nubank	nubank	BANCO
Santander	santander_br	BANCO
Azul	azulinhasaereas	COMPANHIA AÉREA
GOL	voegoloficial	COMPANHIA AÉREA
LATAM	latam_bra	COMPANHIA AÉREA
Avon	avonbr	COSMÉTICOS
Natura	naturabrofficial	COSMÉTICOS
O Boticário	oboticario	COSMÉTICOS
Casas Bahia	casasbahia	E-COMMERCE
Magazine Luiza	magazineluiza	E-COMMERCE
Ponto Frio	pontofrio	E-COMMERCE
Submarino	submarino	E-COMMERCE
CBF	cbf_futebol	ESPORTE
ESPN	espnbrasil	ESPORTE

NBA	nbabrasil	ESPORTE
League of Legends	lolegendsbr	GAMES
Playstation	playstation_br	GAMES
Xbox	xboxbr	GAMES
P&G	pg_brasil	MULTINACIONAL
Unilever	unileverbrasil	MULTINACIONAL
Deezer	deezerbr	STREAMING
Netflix	netflixbrasil	STREAMING
Prime Video	primevideobr	STREAMING
Spotify	spotifybrasil	STREAMING
Acer	acerdobrasil	TECNOLOGIA
Dell	dellnobrasil	TECNOLOGIA
Google	googlebrasil	TECNOLOGIA
Intel	intelbrasil	TECNOLOGIA
Lenovo	lenovobr	TECNOLOGIA
LG	lgdobrasil	TECNOLOGIA
Microsoft	microsoftbr	TECNOLOGIA
Positivo	positivooficial	TECNOLOGIA
Samsung	samsungbrasil	TECNOLOGIA
Claro	ClaroBrasil	TELEFONIA
TIM	TIMBrasil	TELEFONIA
Vivo	vivobr	TELEFONIA
C&A	cea_brasil	VESTUÁRIO
Renner	Lojas_Renner	VESTUÁRIO
Riachuelo	riachuelo	VESTUÁRIO

Fonte: De autoria própria, 2022.

ANEXO B – Script SAS

```
/* BEGINNING PART 1 */
/* ==== EDIT BELOW =====*/

%let project = andre_tcc ;

%let myfolder = &project ;

%let sasusername = u50418389 ;

%let whereisit = /home/&sasusername ; /* online */

libname user "&whereisit/&myfolder";

options fmtsearch=(user work library);

/* END PART 1 */
/* BEGINNING PART 2 */

DATA observed;
  INFILE "&whereisit/&myfolder/&project._counts.txt" ;
  length corpusid $ 7 type $ 11 account $ 15 year $ 4 month $ 3 day $ 3 filenameid $ 8 count
8;
  input corpusid type account year month day filenameid count ;
RUN;

PROC univariate data=observed ;
var count;
run;

data observed;
set observed ;
```

```
timeframe = catx(", year, month, day);  
run;
```

```
data temp (KEEP= type account count);  
set observed;  
run;
```

```
data temp (KEEP=user count);  
set temp;  
x = substr(type, 1, 2);  
user = catx ('_', account, x);  
run;
```

```
ods html file="&whereisit/&myfolder/glm_meta.html";  
ods graphics off;  
proc GLM data=observed;  
    title GLM for dataset = &project._counts TYPE ;  
    class type ;  
    model count = type ;  
    means type ;  
run;
```

```
proc GLM data=temp;  
    title GLM for dataset = &project._counts ACCOUNT ;  
    class user ;  
    model count = user ;  
    means user ;  
run;
```

```
proc GLM data=observed;  
    title GLM for dataset = &project._counts YEAR ;  
    class year ;  
    model count = year ;  
    means year ;  
run;
```

```
proc GLM data=observed;  
    title GLM for dataset = &project._counts QUARTER ;  
    class timeframe ;  
    model count = timeframe ;  
    means timeframe ;  
run;
```

```
ods graphics on;  
ods html close;
```

```
proc sort data=observed;  
by type;  
run;
```

```
proc means data=observed sum mean median;  
var count ;  
class type;  
run;
```

ANEXO C – Relação de usuários

Usuário (_domínio)	Palavras em inglês	Média	Desvio padrão
pierre00_pc	9277	0,54500377	1,13199866
riachuelo_em	46067	0,51509758	0,81771312
playstation_br_em	8616	0,50394615	0,86122817
cea_brasil_em	37085	0,40461103	0,64412020
mmamelotto_pc	14449	0,40141186	0,95546641
azulinhasaereas_em	13516	0,39124001	0,62841360
amazonbr_em	844	0,37796209	0,67353822
positivooficial_em	5067	0,37418591	0,66562775
xboxbr_em	7045	0,36579134	0,79699497
avonbr_em	9216	0,34114583	0,61341819
citroenbrasil_em	921	0,33876221	0,71193718
oficialpeugeot_em	2656	0,32228916	0,63486893
googlebrasil_em	2665	0,32007505	0,61205850
microsoftbr_em	4404	0,31607629	0,67397381
renaultbrasil_em	4642	0,30978027	0,69791081
deezerbr_em	10057	0,30615492	0,54967527
kiadobrasil_em	536	0,29850746	0,56077184
lojas_renner_em	33222	0,29787490	0,59684787
santander_br_em	186467	0,29156902	0,51690490
nissanbrasil_em	2289	0,28877239	0,49389695
pg_brasil_em	726	0,27961433	0,55115965
nubank_em	243572	0,27908791	0,52296244
spotifybrasil_em	2576	0,26708075	0,60555629
micheli_nunes_pc	9044	0,25320655	0,82669153
lollegendsbr_em	3968	0,24873992	0,73369920
vwbrasil_em	3185	0,24427002	0,54679225
acerdobrasil_em	4185	0,21362007	0,48815405
timbrasil_em	46144	0,20475035	0,50606958
vivobr_em	571671	0,20082355	0,44646873
dellnobrasil_em	8199	0,19990243	0,45636106

dhrlo_pc	4920	0,19044715	0,65409574
cortezrafa_in	539	0,18923933	0,47365067
cocacola_br_em	513048	0,18828648	0,51645249
unileverbrasil_em	2452	0,18433931	0,52621162
intelbrasil_em	2430	0,18353909	0,55498102
marcelotas_in	2044	0,17906067	0,63833476
itau_em	123753	0,17813710	0,43437721
tendinick_pc	44416	0,17246037	0,59365389
latam_bra_em	61778	0,17062708	0,41501709
oboticario_em	75758	0,16946065	0,44855552
voegoloficial_em	43958	0,16165431	0,44476221
carlacartas_pc	29252	0,15414331	0,51106414
harloka_pc	1781	0,15103874	0,42424459
bancodobrasil_em	99664	0,14414432	0,39468620
bradesco_em	140624	0,14411480	0,39256809
marxinhu_pc	12235	0,14033510	0,52313211
matheusdevdd_pc	19895	0,13747173	0,49644432
chevroletbrasil_em	14691	0,13702267	0,37585151
cellbit_in	1493	0,13462827	0,43470408
lenovobr_em	8758	0,13450559	0,37592752
tpzanic_pc	28938	0,13439077	0,54504254
prefeitura_cba_go	217	0,13364055	0,37959789
zaczeramano_pc	4754	0,13357173	0,56255667
Oronaldinho_in	486	0,13168724	0,49235813
miumiix_pc	12351	0,13059671	0,49909615
tomcavalcante1_in	2579	0,12989531	0,57941313
tjesoficial_go	1206	0,12935323	0,47669409
cidadepalmas_go	15214	0,12856579	0,37807225
casasbahia_em	116984	0,12812863	0,37186100
vborgaco_pc	7514	0,12456747	0,45994995
manugavassi_in	1643	0,12416312	0,47255411
pitty_in	2216	0,12229242	0,41193194
magazineluiza_em	205577	0,12202241	0,36946528

claudialeitte_in	3396	0,12102473	0,38806147
tacwbellx_pc	25635	0,12085040	0,50977821
gabrielaperes__pc	3395	0,12047128	0,43753180
gouldingagax_pc	91784	0,12043493	0,43199750
rudriek_pc	71207	0,12005842	0,47459724
prefsalvador_go	3844	0,11966701	0,37725924
belbianaa_pc	3699	0,11949175	0,43357740
izareal_in	3222	0,11887027	0,46600440
samguaja_pc	14022	0,11774355	0,43083866
prefeiturapmf_go	6297	0,11386374	0,35942557
luasantana_in	1739	0,11155837	0,38405929
carolhimura_pc	45284	0,11127551	0,46135182
alinebarros_in	1413	0,10969568	0,35304941
victoroliveira_pc	19538	0,10932542	0,44267430
mpf_sp_go	1159	0,10871441	0,33544493
lucasranngel_in	4809	0,10854648	0,38029094
prefmanaus_go	4239	0,10592121	0,34122222
marcosmion_in	4615	0,10574215	0,42269745
paulofilippus_pc	7489	0,10522099	0,42508890
primevideobr_em	32052	0,10517284	0,37313543
scflorianopolis_go	5954	0,10513940	0,34442178
alinecoelha_pc	16287	0,10499171	0,52228163
luuizagoes_pc	12316	0,10498538	0,46608518
tsejusbr_go	11909	0,10269544	0,35413526
netflixbrasil_em	29823	0,10156591	0,38382199
governoes_go	4687	0,10070408	0,34829861
picpay_em	121197	0,10022525	0,32421502
whindersson_in	3776	0,09984110	0,39445155
findbts__pc	40493	0,09833798	0,40999135
pretagil_in	2310	0,09826840	0,34741704
tjrsoficial_go	2239	0,09825815	0,32084507
clarobrasil_em	350825	0,09794912	0,33250595
tjrondonia_go	1185	0,09789030	0,31921135

mcDonalds_br_em	27986	0,09744158	0,34563962
gbyelvalentim_pc	15482	0,09682212	0,38223786
heldermaldonado_pc	19454	0,09679243	0,38615450
dearjooH_pc	49601	0,09677224	0,36126718
maisa_in	29401	0,09628924	0,38749172
tatawerneck_in	5738	0,09532938	0,39801994
samsungbrasil_em	272940	0,09517843	0,33537643
tjgooficial_go	4481	0,09439857	0,31448278
kaliel_pc	21007	0,09358785	0,37816274
fecastanhari_in	3002	0,09327115	0,36412063
lucasluccho_in	1268	0,09305994	0,36524087
mpsp_oficial_go	4210	0,09263658	0,30822715
italopetraglia_pc	14767	0,09243584	0,35250550
rafinhabastos_in	2537	0,09144659	0,36441929
brunabalbino_pc	15483	0,09055093	0,37876113
fab_oficial_go	4619	0,09006278	0,31579397
filiperet_in	2578	0,08999224	0,33379196
defensoriasp_go	1536	0,08984375	0,31633236
naturabroficial_em	12099	0,08934623	0,31865158
tiagosantinelI_pc	8678	0,08896059	0,36235856
ludmilla_in	5026	0,08814166	0,36568024
tjproficial_go	3176	0,08721662	0,31188580
ocaiosan_pc	13777	0,08717428	0,34653416
fabiorchat_in	2666	0,08664666	0,34813069
governosp_go	13745	0,08621317	0,30882883
lucas_in	27009	0,08582324	0,32227380
curitiba_pmc_go	16857	0,08554310	0,31538682
anamariabraga_in	3477	0,08455565	0,32321949
prefeiturabh_go	3111	0,08389585	0,29957268
jessicabatan_pc	20466	0,08340663	0,39817719
paiva_pc	40524	0,08022406	0,32892279
lgdobrasil_em	22421	0,07974667	0,30337185
pontofrio_em	117489	0,07941169	0,30151645

gusttavo_lima_in	2145	0,07738928	0,32406106
tjpe_oficial_go	1897	0,07538218	0,29250514
submarino_em	156077	0,07430948	0,30513850
tjmgoficial_go	2102	0,07278782	0,27584276
lucianohuck_in	1608	0,07213930	0,33803841
danilogentili_in	10738	0,07180108	0,35532923
fepaesleme_in	7004	0,07167333	0,30977690
t3ddyyyyy_in	3040	0,07072368	0,29462354
igorafirma_pc	11341	0,07071687	0,35601224
govparaiba_go	9703	0,07039060	0,29545503
shadesofbru_pc	11392	0,06855688	0,32123471
thiaguinhocomth_in	2239	0,06699419	0,29438510
pedrovieira_25_pc	23147	0,06575366	0,31094572
luisasonza_in	7140	0,06554622	0,30019006
tjmaoficial_go	2277	0,06543698	0,26945389
brunogagliasso_in	9741	0,06446977	0,27887884
ivetesangalo_in	1650	0,06424242	0,27553389
paulafernandes7_in	921	0,06406080	0,26214088
prefeitura_rio_go	17707	0,06364715	0,26816369
prefeitura_poa_go	18606	0,06347415	0,27936630
loneliestgrl_pc	28709	0,06290710	0,36259508
tjpaoficial_go	1949	0,06259620	0,24857217
governodoamapa_go	6574	0,06191056	0,25984419
dpuoficial_go	4580	0,06157205	0,25708424
alanzikacrvg_pc	1884	0,06104034	0,30045139
jose_simao_in	1988	0,06086519	0,25542460
tjspoficial_go	4627	0,06073049	0,25800405
tjmtoficial_go	3403	0,06053482	0,27299040
sabrinasato_in	5837	0,05910570	0,26200070
micheltelo_in	1395	0,05878136	0,26941129
babi_pc	67234	0,05825921	0,27775280
jorgeemateus_in	4690	0,05799574	0,26051164
alvxaro_in	13685	0,05648520	0,27239464

julianapaes_in	3057	0,05561008	0,25871352
tjscoficial_go	3954	0,05538695	0,24271332
dimitrisanders_pc	50689	0,05320681	0,24911068
stf_oficial_go	6171	0,05250365	0,23577481
gessicakayane_in	5621	0,05194805	0,26303430
anaclaralimaz_pc	1426	0,05189341	0,23419873
prefeiturabelem_go	9529	0,05110715	0,23092893
natalprefeitura_go	10703	0,05101373	0,24760986
stjnoticias_go	7835	0,05003191	0,23165043
cnj_oficial_go	12561	0,04959796	0,22994400
mpf_pgr_go	11611	0,04917750	0,22976595
edujbatista_pc	14868	0,04694646	0,22056010
guarana_em	44990	0,04661036	0,23607302
oserginho_in	1141	0,04645048	0,29990953
brunorahall_pc	3172	0,04634300	0,25251031
prefmaceio_go	3806	0,04519180	0,21398356
dedesecco_in	3522	0,04457694	0,22735493
marmilbr_go	4181	0,04448696	0,25589949
rdaedua_pc	2818	0,04364798	0,23655264
prefdeteresina_go	5859	0,04352279	0,22398883
tjtocantins_go	4070	0,04324324	0,21744353
exercitooficial_go	7162	0,04286512	0,27558076
tjrjoficial_go	1476	0,04268293	0,20880771
marceloadnet_in	9451	0,04264099	0,26423808
gio_antonelli_in	1878	0,03940362	0,22028718
fiuk_in	6323	0,03827297	0,22524125
gabigol_in	1350	0,03777778	0,29206625
wesleysafadao_in	6151	0,03722972	0,19691705
bmcchd_pc	12276	0,03633105	0,20931395
prefaracaju_go	16031	0,03555611	0,19503075
tjacoficial_go	4005	0,03445693	0,19306466
tjdftoficial_go	3399	0,03353928	0,18648930
senadofederal_go	22840	0,03121716	0,18202684

zamenza_pc	155908	0,02999846	0,17776454
pefabiodemelo_in	2470	0,02995951	0,18639749
camaradeputados_go	30181	0,02369040	0,15869554
pekenozapp_pc	5804	0,02033081	0,15173246
tjamazonas_go	688	0,01889535	0,13625454
bryasant__pc	15225	0,01490969	0,14028710