UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
FACULDADE DE ARTES COMUNICAÇÃO E LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO
EM ESTUDOS DE LINGUAGENS
CURSO DE MESTRADO

MARINA LUZ

FRASEOLOGIA E INFORMÁTICA: CONSTITUIÇÃO DE *CORPUS* DE COMENTÁRIOS DO TWITTER PARA O ESTUDO DE EXPRESSÕES IDIOMÁTICAS

Campo Grande – MS
Fevereiro – 2022

MARINA LUZ

FRASEOLOGIA E INFORMÁTICA: CONSTITUIÇÃO DE *CORPUS* DE COMENTÁRIOS DO TWITTER PARA O ESTUDO DE EXPRESSÕES IDIOMÁTICAS

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre ao Programa de Pós-Graduação em Estudos de Linguagens da Faculdade de Artes, Letras e Comunicação da Universidade Federal de Mato Grosso do Sul, sob a orientação da Professora Drª Elizabete Aparecida Marques e coorientação do Profº Drº Fabrice Charles Bernard Isaac.

Área de Concentração: Linguística e Semiótica.

Campo Grande – MS

Fevereiro – 2022

MARINA LUZ

FRASEOLOGIA E INFORMÁTICA: CONSTITUIÇÃO DE *CORPUS* DE COMENTÁRIOS DO TWITTER PARA O ESTUDO DE EXPRESSÕES IDIOMÁTICAS

MEMBROS COMPONENTES DA BANCA EXAMINADORA

Presidente e Orientadora: Profª Drª Elizabete Aparecida Marques

Universidade Federal de Mato Grosso do Sul, Campus de Campo Grande.

Coorientador: Profº Drº Fabrice Charles Bernard Isaac

Université Sorbonne Paris Nord, Paris, França

Membro Titular: Profª Drª Aparecida Negri Isquerdo

Universidade Federal de Mato Grosso do Sul, Campus de Campo Grande.

Membro Titular: Profª Drª Angela Karina Manfio

Universidade Estadual de Mato Grosso do Sul, Campus de Dourados.

Data da defesa: 04/02/2022

Campo Grande – MS, 04 de fevereiro de 2022.

À minha família por todo amor e companheirismo.

AGRADECIMENTOS

Aos meus pais João e Mirian, minha cunhada Juliana, e minha sobrinha Maria, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho. Ao meu irmão Diogo, que desde quando éramos crianças me inspirou com sua curiosidade e sua paixão pelo mundo da informática, do *marketing* e da programação de computadores.

Ao meu namorado Arthur, pelas longas conversas, pelo companheirismo e pela cumplicidade em todos os momentos que partilhamos.

Aos amigos, que sempre estiveram ao meu lado, pelo carinho incondicional e pelo suporte demonstrado ao longo de todo o período em que me dediquei a esta pesquisa. Elizaveta, Erika, Luiz, Mariana, Natália, Pedro, Quentin e Rodrigo, amo vocês!

À minha orientadora Elizabete Aparecida Marques e meu coorientador Fabrice Charles Bernard Isaac, que me inspiraram e conduziram esta pesquisa com paciência e dedicação, sempre disponíveis a compartilhar todo o seu vasto conhecimento. Obrigada! *Merci!*

À Profª Drª Aparecida Negri Isquerdo e ao Prof. Dr. Renato Rodrigues Pereira, pelas considerações feitas no processo de qualificação. Elas foram fundamentais para delinear os caminhos deste trabalho.

Aos meus colegas de trabalho da Escola Estadual Maria de Lourdes Toledo Areias, Adriana, Edi Carlos, Heraldo, Maria Eva e Tays. Obrigada pelo incentivo e pela amizade!

À CAPEs pelos auxílios financeiros oferecidos durante seis meses desse processo, auxiliando de forma considerável na minha permanência no programa de mestrado.

A todos que participaram, direta ou indiretamente ao longo desses dois anos de estudo, enriquecendo o meu processo de aprendizado e construindo memórias afetivas, dispenso minha eterna gratidão.

“Una lengua es más que un conjunto de categorías fonológicas, morfológicas, sintácticas o léxicas y una serie de reglas para su uso. Una lengua, existe en el contexto de prácticas culturales que, a su vez, descansan en algunos recursos semióticos, como las representaciones y expectativas que proporcionan los cuerpos y movimientos de los participantes en el espacio, el entorno construido en el que interactúan, y las relaciones dinámicas que se establecen por medio de la recurrencia en la actividad conjunta que realizan.” (DURANTI. 2000, p. 104).

SUMÁRIO

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

RESUMO

Desde as primeiras investigações, os estudos do léxico voltaram-se para as necessidades comunicativas do homem. Considerando, portanto, que os contextos e propósitos comunicativos interferem na seleção léxica dos indivíduos, reitera-se a relação indissociável e contínua entre língua, sociedade e cultura (BARBOSA, 1987, p. 158), na qual, a primeira desenha-se como um fenômeno social de "natureza dinâmica" (BIDERMAN, 1992, p.32) capaz de perpetuar as heranças culturais de um povo. Esse processo de comunicação é cotidianamente potencializado diante das interações propostas pelas redes sociais, ambientes virtuais de confraternização que permitem uma manifestação espontânea e despojada da linguagem, especificidades essas que compartilham o caráter de coloquialidade das unidades fraseológicas da língua, sobretudo as Expressões Idiomáticas (EI). Objeto de estudo deste trabalho, as EI definem-se como estruturas léxicas complexas, plurilexicais, indecomponíveis, de sentido figurado e cristalizadas pela tradição cultural (XATARA, 1998; 2013). Nesse cenário, desenha-se esta pesquisa que, em uma perspectiva geral, objetiva a elaboração de um *corpus* linguístico de expressões idiomáticas, por meio de um banco de dados de comentários coletados da rede social *Twitter*. Aliado a esse propósito, de forma específica, realiza-se a elaboração de códigos computacionais que auxiliem as ferramentas para a coleta dos textos que compõe a base de dados, do mesmo modo que, promove-se uma discussão acerca da produtividade de EI, considerando a hipótese de que os usuários que utilizam o *Twitter* empregam o uso de uma linguagem mais coloquial. A investigação pauta-se nas bases teóricas da Lexicologia, da Fraseologia, da Linguística Computacional e da Linguística de *Corpus*. A metodologia utilizada nesta pesquisa tem caráter descritivo e quantitativo. Os resultados alcançados montam uma base de dados de coletas diárias de *tweets*, no período de 6 meses, dezembro de 2020 à maio de 2021, da mesma forma que uma amostra contendo 232 EI, sendo 50% de caráter somático. Além disso, mostraram-se mais produtivas as *hashtags* "#" com temática relacionada à *reality shows* e política.

Palavras-chave: Fraseologia; Expressões Idiomáticas; *Twitter*; Linguística Computacional; Linguística de *Corpus*.

RESUMEN

Desde las primeras investigaciones, los estudios del léxico se han vuelto hacia las necesidades comunicativas del hombre. Teniendo en cuenta que los contextos y propósitos de comunicación interfieren en la selección léxica de los individuos, se reitera la relación inseparable y continua entre lengua, sociedad y cultura (BARBOSA, 1987, p. 158), en el que primero está diseñado como un fenómeno social de "naturaleza dinámica" (BIDERMAN, 1992, p.32) capaz de perpetuar la herencia cultural de un pueblo. Ese proceso comunicativo se ve potenciado día a día por las interacciones que proponen las redes sociales, entornos virtuales de socialización que permiten una expresión espontánea y despejada del lenguaje, especialmente las locuciones. Objeto de estudio de este trabajo, las locuciones se definen como estructuras léxicas complejas, plurilexicales, indecomponibles, con sentido figurativo y cristalizadas por la tradición cultural (XATARA, 1998). En este escenario, se diseña esta investigación que, en una perspectiva general tiene como objetivo la elaboración de un corpus lingüístico de locuciones, a través de una base de datos de comentarios recolectados de la red social Twitter. Aliado a este propósito, en específico, se lleva a cabo la elaboración de códigos de computacionales para ayudar a las herramientas de software que recolectan los textos de la base de datos, así como una discusión considerando la hipótesis de que los usuarios de Twitter empleen el uso de un lenguaje más coloquial. La investigación se sustenta en las bases teóricas de la Lexicología, Fraseología, Lingüística Computacional y Lingüística de *Corpus*. La metodología utilizada en esta investigación es descriptiva y cuantitativa. Los resultados obtenidos ensamblan una base de datos de recopilaciones diarias de tweets, durante un período de 6 meses, de diciembre de 2020 a mayo de 2021, así como una muestra que contiene 232 locuciones, las cuáles, el 50% son somáticas. Además, los hashtags "#" con temas relacionados con reality shows y política fueron más productivos.

Palabras-clave: Fraseología; Locuciones; Twitter; Lingüística Computacional; Lingüística de *Corpus*.

INTRODUÇÃO

Desde os tempos antigos, a linguagem é concebida como uma capacidade humana que intriga e interessa sobremaneira aos filósofos e demais estudiosos, especialmente os linguistas, de distintos países do mundo. As primeiras investigações acerca dela remontam aos hindus no século IV a.C. e percorrem historicamente os séculos através das pesquisas de Platão[1], Varrão[2], Franz Bopp[3], Saussure[4], Chomsky[5] e muitos outros (PETTER, 2002).

Em inglês, o termo "linguagem" pode ser traduzido como *language*, o qual por sua vez também denota "língua". O duplo significado tem referência na etimologia do idioma, mas pode ser associado informalmente à relação intrínseca que ambos termos carregam. Dessa forma, língua e linguagem não possuem uma mesma significação, mas mantêm uma conexão ativa e veemente, ao ponto de que para definir linguagem é necessário, impreterivelmente, voltar a atenção para o conceito de língua.

Saussure (1969) entende a linguagem como o conjunto formado por dois componentes: *langue* e *parole* – língua e fala. Nessa perspectiva, a língua é definida pelo autor como um sistema de signos, ou seja, um conjunto de unidades que se relacionam de modo ordenado dentro de um todo. Por outro lado, a fala é, então, o ato individual resultado das combinações feitas pelo sujeito falante ao utilizar os códigos da língua.

Entretanto, o conjunto linguagem-língua-fala vai além de possibilitar que o homem nomeie os elementos do mundo e se expresse enquanto indivíduo. O ser humano, como um sujeito integrante da vida em sociedade, carrega de forma inerente a necessidade de comunicar e interagir. Segundo Bakhtin[6] e Voloshinov[7] (1988), língua, sociedade e sujeito contam com uma condição de interligação intrínseca, de modo que as formas de interação social entre os indivíduos moldam a língua. Isso é dizer que a relação de comunicação entre sujeito e mundo é, na verdade, uma via de mão dupla. A língua seria então um fenômeno social que permite a comunicação entre os falantes.

---

[1] Platão (428/427 a.C. – 348/347 a.C.) foi um filósofo e matemático do período clássico da Grécia Antiga.
[2] Marco Terêncio Varrão (116 a.C. - 27 a.C.) foi um filósofo e antiquário romano de expressão latina.
[3] Franz Bopp (1791 – 1867) foi um linguista alemão e professor de filologia e sânscrito na Universidade de Berlim.
[4] Ferdinand de Saussure (1857 - 1913) foi um linguista e filósofo suíço, cujas elaborações teóricas propiciaram o desenvolvimento da linguística enquanto ciência autônoma.
[5] Avram Noam Chomsky (1928) é um linguista, filósofo, sociólogo, cientista cognitivo, comentarista e ativista político norte-americano, considerado por muitos acadêmicos como o "o pai da linguística moderna".
[6] Mikhail Mikhailovich Bakhtin (1895 - 1975) foi um filósofo e pensador russo.
[7] Valentin Nikoláievitch Voloshinov (1895 - 1936) foi um filósofo, músico, linguista e crítico literário russo ligado ao Círculo de Mikhail Bakhtin.

Segundo o dicionário online, Michaelis[8], o verbo "comunicar" tem sua origem no latim (*communicare*) e significa "1. Transmitir conhecimento, informação, mensagem, etc.; participar, significar. 2. Pôr em contato ou ligação". Nessa perspectiva, infere-se que para que haja comunicação, ou seja, transmissão de conhecimento, participação sócio interativa, faz-se necessária essa relação dualística – língua e linguagem, inicialmente descrita por Saussure, que é capaz de influir nos mais diversos meios e contextos e, por sua vez prover a relação indivíduo e sociedade.

Segundo Barbosa (1991), "língua, sociedade e cultura são indissociáveis, interagem continuamente, constituem, na verdade, um único processo complexo [...]", haja vista, que são os propósitos comunicativos/contextos que interferem nas atividades discursivas, perpassando pelas intenções dos indivíduos e pelos espaços sociais, determinando por suas escolhas lexicais do falante. Parafraseando Biderman (1978), os Estudos Lexicais - melhor tratados no capítulo 1 - consideram a língua como um instrumento de comunicação de um povo específico, e os aspectos históricos, sociais e culturais estão inter-relacionados. Frente à essas ponderações, nesta pesquisa, considera-se a língua enquanto um fenômeno social e não como um mero rotulador de objetos e conceitos.

A comunicação humana é diariamente potencializada pelas novas possibilidades de interação social promovidas pelos facilitadores eletrônicos e atravessadas pelo sistema global de redes amplamente conhecido como internet. Dentro desse universo, destacam-se as redes sociais[9], as quais, possibilitam que grandes bases de usuários possam publicar vídeos e fotos pessoais e até mesmo realizar divulgações profissionais, ou ainda discutir temas comuns em grupos públicos e/ou privados, juntamente com outros usuários que partilhem dos mesmos interesses.

Pesquisadores, como Silva (2011) por exemplo, afirmam que essas plataformas surgiram como evolução dos fóruns virtuais, nos quais várias pessoas podiam se comunicar sobre determinado assunto que lhes interessavam. Ademais, a história aponta que a primeira rede social, em formato que se assemelha aos atuais, foi a norte-americana *Six Degrees* criada por Andrew Weinreich em 1997, entretanto, somente a partir dos anos 2000 os avanços desses

[8] https://michaelis.uol.com.br/

[9] Musso (2006) define rede social como "uma das formas de representação dos relacionamentos afetivos, interações profissionais dos seres humanos entre si ou entre seus agrupamentos de interesses mútuos." Nesse ínterim, os estudos de Castells (1999, 2003, 2009, 2011) apresentam as redes sociais como uma organização social localizada no espaço virtual capaz de modificar a relação do indivíduo com o mundo, possibilitando novos olhares diante da experiência, do poder e da cultura.

recursos sociais começaram a se tornar expressivos. A partir de então, uma enxurrada de plataformas sociais surgiu e, atualmente, algumas das mais usadas têm uma quantidade de usuários que ultrapassa a população de muitos países.

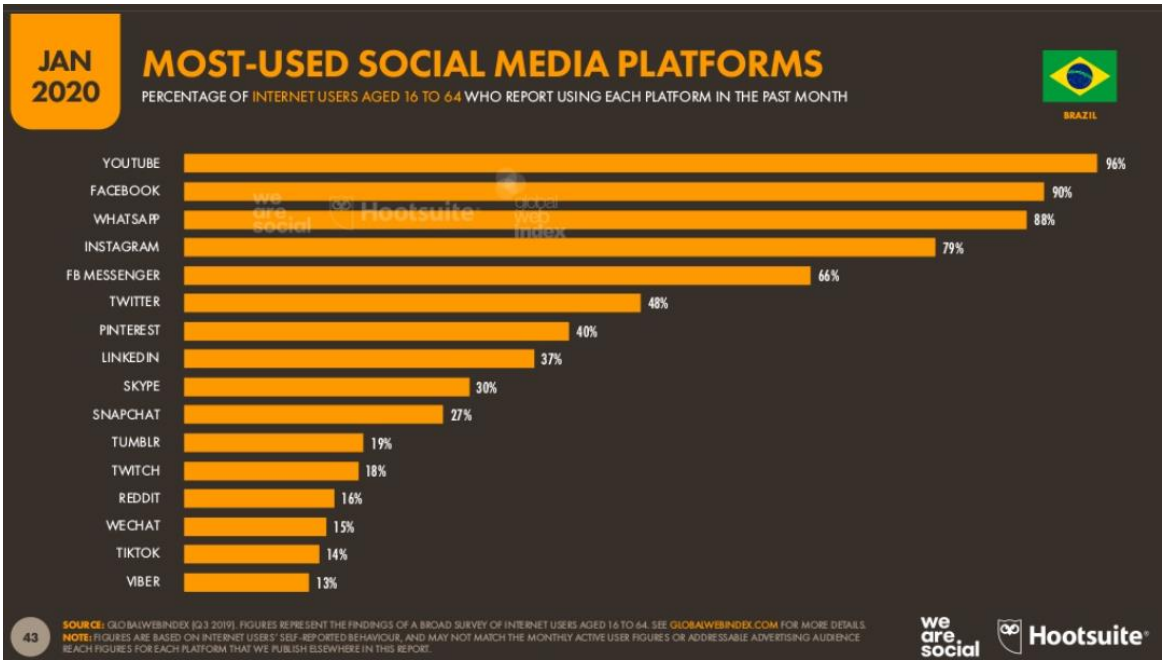
Respeitadas agências de *marketing*[10] digital realizam pesquisas mundiais sobre a atuação e presença das mídias sociais no cotidiano da população. A união de duas dessas empresas, *Hootsuite* - sistema norte-americano especializado em gestão de marcas – e *We are social* - agência britânica de mídia social -, consolidou um relatório anual de mais de 200 páginas - *Global Digital Reporter* - como um dos materiais mais aguardados por gestores de *marketing* do mundo todo.

Contendo informações essenciais e *insights*[11], como crescimentos de cada rede social, características demográficas relacionadas ao uso da internet e as principais tendências de *e-commerce*[12], a exposição das agências dita a forma como milhares de marcas irão se movimentar economicamente no ano seguinte.

De acordo com o relatório de 2020[13] disponibilizado pela *We are social* e *Hootsuite*, 96% da população brasileira é usuária frequente da internet e 84% está conectada e contribui em continuidade linear com as redes sociais. Esse valor considera aproximadamente e respectivamente 201 e 175 milhões de pessoas ativas virtualmente de um total de 209,5 milhões de brasileiros. O estudo aponta, ainda, as plataformas sociais públicas e privadas mais usadas, e a porcentagem estimada de usuários presentes em cada plataforma social. As informações comentadas podem ser conferidas na sequência, por meio da figura 1.

[10] Segundo o economista Philip Kotler (1999), "*marketing* é a ciência e arte de explorar, criar e entregar valor para satisfazer as necessidades de um público-alvo com geração de lucro". Em outras palavras, o objetivo geral do *marketing* é criar valor e chamar a atenção do cliente, gerando lucros tanto para vendedor quanto para o comprador.
[11] Compreensão de uma causa e efeito específicos dentro de um contexto particular.
[12] Transação comercial realizada através de equipamento eletrônico, como, por exemplo, computador, *tablet* e *smartphone*.
[13] Disponível em: https://datareportal.com/reports/digital-2020-brazil. Acesso em 26 de março de 2021.

Figura 1: As plataformas de mídias sociais mais usadas no Brasil em 2020.



Fonte: *We are social – Global Digital Reporter*, 2020.

Entende-se que as redes sociais, além das funções principais a que se propõem, podem ser entendidas como grandes vitrines do comportamento humano. Atualmente, inclusive, não é esporádico encontrar essas plataformas como base para diversos estudos nas mais distintas áreas, como por exemplo aqueles com viés antropológico, ou ainda político, que analisam o discurso, a atuação cultural, e principalmente a influência desses instrumentos no cotidiano da sociedade. Se as redes envolvem conduta humana, é natural concluir que circundam, também, a língua e suas alterações, de modo que se apresentam como campo para investigações lexicais.

Nesse âmbito, é costumeiro observar o desenvolvimento de estudos lexicais que utilizam *corpus* linguístico[14] pré-constituído[15] com material diacrônico, histórico e contemporâneo, utilizando como fonte, em sua maioria, textos jornalísticos, pesquisas e entrevistas. Seria orgânico pensar, então, que por carregar a característica espontânea que os *corpora* requisitam, as redes sociais são esferas apropriadas para a constituição desses conjuntos representativos da linguagem humana.

Entretanto, é circunstancial, ainda, encontrar esses textos como elementos principais na formação de *corpus.* O fato pode ser justificado, talvez, pela "pouca idade" que as redes sociais

[14] *Corpus* linguístico pode ser entendido como um conjunto de registros espontâneos orais ou escritos de textos, em uma língua específica, que servem como análise para estudos sobre comportamento linguístico. Pl. *corpora*.
[15] É possível elencar uma lista de *corpora* prontos para uso, como por exemplo: Banco do Português; CetenFOLHA; CetenPUBLICO; COMPARA; *Corpus* do Português; *Corpus* Brasileiro; Lácio-Web; Linguateca; PortPopular; PHPB; TychoBrahe.

têm quando comparadas às mídias jornalísticas, por exemplo. Porém, a questão é que a língua nesses ambientes é manifestada com muita naturalidade e pode ser um campo rico no que diz respeito às especificidades linguísticas que compartilham desse despojamento no nível discursivo, como é o caso unidades fraseológicas (UF).

As UF são estruturas léxicas complexas da língua e que apresentam seu conteúdo formado por duas ou mais unidades lexicais, de caráter relativamente estável, tais combinações apresentam certo grau de idiomaticidade. Comumente conhecidas como frases feitas, ditos populares ou formas de dizer, as UF refletem de maneira ativa a relação: léxico – cultura. Nesse grupo destacam-se as expressões idiomáticas, os provérbios, os clichês, as colocações, dentre outros.

Nesse cenário, seleciona-se como objeto de estudo desta pesquisa, as Expressões Idiomáticas (EI) que de forma breve, definem-se como "lexias complexas e indecomponíveis, conotativas e cristalizadas em um idioma pela tradição cultural" (XATARA, 1998). Refere-se, portanto, àquele grupo de palavras que não podem ser separadas nem ter posição alterada, que tem sentido figurado, isto é, tem sentido diferente do apresentado literalmente, e que foram consolidadas pela comunidade que as usa.

A termos de exemplificação tem-se a EI *lavar as mãos,* que é composta por mais de uma unidade léxica (lavar + as + mãos) e, não podendo sofrer alteração na ordem, substituição ou supressão das lexias (**mãos lavar*, **lavar os braços*, **limpar as mãos*, **lavar*, **mãos*) – indecomponível - não significando, ao ser aplicada em um discurso, literalmente a ação de lavar as mãos, mas sim a atitude de não se envolver em algo ou alguma situação – conotativa - e tendo significado amplamente reconhecido pela sociedade brasileira – cristalizada.

As EI podem ser encontradas nas mais distintas situações sociais e têm lugar quase que cativo nos contextos caracterizados por um caráter informal e descontraído. Dentre tantos cenários, um local que compreende tais atributos e merece atenção especial, como já dito anteriormente, é a esfera das redes sociais. Isto se dá porque, além das propriedades já citadas, a plataforma social é um tipo de mídia que permite a comunicação totalmente espontânea do usuário. Assim, a produção de textos não é controlada e, portanto, gera a oportunidade de coleta e armazenamento desse material para análises que preconizem essa naturalidade.

Retornando a atenção para a figura 1, é possível verificar que a citada pesquisa não é voltada somente para as redes públicas, mas também para as de conteúdo privado como *WhatsApp* e *Messenger*. Para a presente discussão, as redes privadas não contemplam as características necessárias justamente pelo fato de ser impossível coletar textos de terceiros sem

autorização prévia. Além disso, também não seria viável trabalhar neste momento com todas as redes sociais públicas, visto que o volume de material precisaria de um tempo maior para análise e tratamento adequados.

Assim, para a seleção de uma rede social, foram inicialmente elencadas as seis públicas mais ativas: *YouTube, Facebook, Instagram, Twitter, Pinterest* e *Linkedin*. Dentre elas, preliminarmente exclui-se a rede *Linkedin,* isto porque seu objetivo principal é conectar pessoas através de suas identidades profissionais buscando propiciar vínculos acadêmicos e empregatícios. Tal realidade está intrinsecamente relacionada à um carácter formal e pouco espontâneo, aspectos que não combinam com as características comentadas das expressões idiomáticas.

Do mesmo modo, é exclusa a rede *Pinterest*, por ser uma plataforma de compartilhamento de fotos que se assemelha a um quadro de inspirações. Nesse espaço, os usuários se expressam por meio do gerenciamento de imagens, e textos imagéticos, que embora possam representar expressões idiomáticas, não contemplam o tipo de pesquisa lexical aqui desejada.

Por outro lado, *Facebook* e *Instagram* são mídias sociais criadas em 2004 e 2010, respectivamente, onde os usuários cadastrados podem conectar-se com outros indivíduos, independente do seu país de residência, para compartilhar as informações que forem convenientes. Assim, cada sujeito com seu "Mural" publica comentários, fotos e vídeos, e outras pessoas podem comentar esses compartilhamentos.

Embora sejam ambos espaços informais de produção espontânea, para acessar as duas redes o usuário precisa criar uma conta e seu perfil midiático só se torna visível aos demais se ele assim o desejar. Essa possibilidade de restrição limita também a coleta de dados para a pesquisa aqui intencionada e, portanto, tira as duas mídias da lista de candidatos.

O *YouTube*, por sua vez, é uma plataforma norte-americana de compartilhamento de vídeos, criada em 2005 e atualmente subsidiária da empresa *Google*, que pode ser acessada em dez idiomas, são eles: inglês, espanhol, português, francês, árabe, russo, hindi, malaio e bengalês. Nessa rede, os usuários, a partir do momento em que realizam um cadastro, estão aptos para publicar vídeos, criar listas de reprodução, avaliar publicações e comentar em um espaço existente em cada vídeo.

Ainda que seja preciso um cadastro para realizar as possibilidades listadas, as informações publicadas permanecem visíveis para todos que tiverem interesse e, por conseguinte, é um bom candidato à fonte da investigação aqui comentada. Entretanto, no caso

do *YouTube* a principal matriz de material para a pesquisa seriam os comentários, e estes são enviados pelos usuários para compartilhar opiniões sobre assuntos específicos tratados em vídeos. Dessa maneira, para coletar os textos seria necessário escolher estilos e/ou assuntos representativos para selecionar vídeos determinados. Tal triagem e manipulação podem tornar o conteúdo avaliado menos espontâneo e, assim, comprometer o resultado da pesquisa.

Resta, portanto, plataforma de *microblogging Twitter* criada em 2006. A rede funciona de modo similar ao *Facebook* e *Instagram*, porém enquanto os dois são voltados principalmente para fotos e vídeos, o *Twitter* preconiza comentários escritos sobre o cotidiano e conta com a alternativa de acessar um espaço denominado como *Trending Topics.*

Esse ambiente dentro da plataforma permite que os usuários acessem os assuntos mais comentados do dia, e, assim, possibilita que determinada pesquisa balize a coleta de dados através da delimitação espontânea dos próprios usuários. Isto é, com o *Twitter* não é necessário escolher um assunto e assim influenciar na naturalidade da coleta de material, os assuntos são elencados pela quantidade de comentários originados a partir deles.

São notáveis os números de pesquisas e artigos que se propõem a estudar as redes sociais em suas mais diversas particularidades. Entretanto, uma busca detalhada mostra a vacância de *corpora* formados por textos em língua portuguesa provenientes de redes sociais, de modo que irrompe a possibilidade de constituição de um *corpus* dentro dos parâmetros citados.

Além disso, é também perceptível a escassez de trabalhos que se propõem a estudar as plataformas virtuais alinhadas ao estudo das expressões idiomáticas. Desse modo, destaca-se a relevância natural em trazer esses objetos para o foco das discussões promovendo, então, a percepção de que a língua como objeto vivo percorre os mais diferentes canais de projeção e, portanto, a Linguística[16] encontra campo para análise.

Diante das considerações elencadas, nesta dissertação, de forma ampla, objetiva-se a elaboração de um *corpus* linguístico, constituído por meio de comentários, sobre os assuntos mais debatidos, em período específico, na rede social *Twitter*, verificando ferramentas informáticas que auxiliem nessa composição. Utilizando-se das bases teórico-metodológicas da Lexicologia, da Fraseologia e da Linguística de *Corpus* pretende-se, ainda, de maneira específica:

- Elaborar códigos e matrizes computacionais que funcionem como ferramentas virtuais para a coleta diária dos comentários dispostos nos *Trending Topics*.

[16] Área do conhecimento responsável por estudar a linguagem verbal humana a partir de teorias que analisam a evolução e desdobramentos das línguas. Além de aspectos gramaticais, como aqueles que envolvem a estrutura dos idiomas, a Linguística também pode ocupar-se de tópicos sociais implicados nas variações da linguagem.

- Promover a discussão sobre a produtividade de um *corpus* desse tipo, no que diz respeito à uma amostra de ocorrência de expressões idiomáticas, considerando a hipótese de que os usuários, no geral, se comunicam por meio de uma linguagem mais coloquial.

- Contribuir com o processo de consolidação dos estudos no âmbito da Lexicologia e da Fraseologia, sobretudo na relação que essas disciplinas traçam com a Informática.

Para atingir tal objetivo, esta dissertação estrutura-se em 3 capítulos. **No primeiro capítulo**, desenvolve-se as bases teóricas que englobam as premissas essenciais da Lexicologia (mantem-se o foco no conceito de léxico), da Fraseologia (discorre-se sobre a história e desenvolvimento da Teoria Fraseológica, sobretudo no que diz respeito aos conceitos das UF e das EI) e da Linguística Computacional (pontua-se o contexto histórico e sua relação com a Linguística de Corpus).

No segundo capítulo, reserva-se a atenção aos procedimentos metodológicos de constituição de *corpus* e análise do material, e apresenta informações sobre: a fonte do *corpus*, ou seja, a rede *Twitter*; o modo como o *corpus* foi constituído, e, assim, tópicos específicos relacionados à Linguística computacional e; os procedimentos de análise adotados para o *corpus* já formado.

Por fim, **o terceiro capítulo** apresenta as discussões acerca do *corpus* gerado e o tratamento das Expressões Idiomáticas encontradas, visando a confirmação da hipótese de que há uma grande produtividade de EI nesse espaço virtual, visto que, os usuários tendem a utilizar uma linguagem mais coloquial e espontânea.

Ademais, este trabalho conta com a presença de **elementos pré e pós-textuais** que têm por objetivo organizar e facilitar a leitura e compreensão do mesmo. Tais elementos são: Resumo, Lista de Figuras e Quadros, Introdução, Considerações Finais, Referências Bibliográficas e Anexos.

CAPÍTULO 1 – OS ESTUDOS DO LÉXICO E AS EXPRESSÕES IDIOMÁTICAS

Este capítulo tem como objetivo principal discutir os fundamentos teóricos que embasaram a pesquisa. Considerando que as Expressões Idiomáticas correspondem a uma parte significativa do sistema lexical da Língua Portuguesa, busca-se centrar a investigação no âmbito da Lexicologia, apresentando e clarificando a conceituação de termos substanciais como "léxico", "lexema", "lexia" e "lema". Além disso, é concebido um percurso histórico que evidencia tanto de que maneira a Fraseologia passou a ser entendida como uma disciplina, quanto quais são os elementos estudados por ela. Nessa perspectiva, ademais de explicitar definições e características de Enunciados Fraseológicos, Colocações e Locuções, reserva-se um espaço específico para discussões teóricas e práticas sobre as expressões idiomáticas.

Paralelamente, destina-se espaço, também, para dissertar sobre Linguística Computacional, outro pilar fundamental do estudo. Nesse âmbito, são exibidas, de maneira diacrônica, a origem e evolução do computador, e comentados aspectos intrínsecos a ele, como a "computabilidade".

Na sequência, são expostos e discutidos preceitos teóricos da Linguística baseada em *corpus*, sua relação com a Fraseologia e, especificamente, com as expressões idiomáticas. Por fim, a discussão é direcionada para o processamento de linguagem natural e, portanto, as possibilidades relacionadas ao tratamento automático da língua, no que tange aos objetivos da pesquisa aqui retratada.

1.1 LEXICOLOGIA

> A língua é o alicerce da civilização. É a cola que une as pessoas. (A CHEGADA, 2016).

As primeiras averiguações acerca da palavra têm registro datado no período da Antiguidade Clássica, no entanto, tais investigações não se detinham nos estudos lexicais em si, mas tinham foco nos âmbitos fonológicos, morfológicos e sintáticos. Foi somente na Idade Média, sobretudo nos séculos XVI a XVIII que devido as necessidades econômicas, sociais e culturais com a expansão mercantil, ressaltou a necessidade de descrição do léxico. Contudo, é somente por volta do XIX que essas reflexões passam a ter cunho e rigor científico, culminando por fim, no nascimento da Lexicologia, enquanto uma disciplina voltada ao estudo do léxico.

Nos anos seguintes a esse período, a Lexicologia fortalece seus estudos, solidificando seus conceitos e teorias. Assim, estabelece-se, aos poucos uma abordagem epistemológica

metalexicográfica, que tem como objetivo a análise e descrição lexical. Tal disciplina, interessa-se em discutir o signo linguístico na sua totalidade, como concebido por Saussure (2006) e toma como cerne inicial a definição do conceito de palavra. Em suas reflexões a Lexicologia destina-se a explicar o processo de formação das palavras e da etimologia relacionando, na medida em que se faça necessário, com os estudos fonéticos, fonológicos, morfológicos, semânticos, dentre outros.

Em concordância com o posicionamento de Barbosa (1991, p. 53), nesta pesquisa considera-se a Lexicologia como "um dos ramos da Linguística" que objetiva a análise e descrição do léxico. De forma mais específica, Biderman (2001, p. 16) reitera que a disciplina é "ciência antiga que tem como objetivos básicos de estudo e análise, a palavra, a categorização lexical e a estruturação do léxico". Dessa forma, pode-se dizer que ela considera, portanto, o estudo das unidades lexicais de uma ou várias línguas, isoladas ou em contato, mediante todos os seus aspectos, considerando as características que vão desde o nível fonético ao discursivo.

Destaca-se por fim, por entender como relevante para a discussão dos resultados desta dissertação, a teoria da Lexicologia Social de Matoré (1953). Algumas das ideias essenciais do pressuposto citam que: (i) a criação de uma nova palavra é assimilada como a construção de um conceito; (ii) as palavras estão presentes na consciência dos falantes em relações recíprocas - no contexto - e associativas - fora do contexto -; (iii) a palavra é considerada como um reflexo da comunidade que a utiliza, logo carrega caráter social; (iv) as palavras se organizam em conjuntos - campos nocionais -; (v) cada campo nocional é composto por novas palavras - palavras-testemunho -, essas, por sua vez, entendidas como sinais de uma nova situação social; (vi) todo campo nocional possui como centro uma palavra-chave, que reflete o ideal da sociedade da época em questão.

Ademais, é importante ater-se ao seu objeto de estudo da Lexicologia: o léxico. Do mesmo modo que, faz-se necessária a distinção de termos que são fundamentais dentro dos estudos lexicológicos, e que por vezes pode gerar certa confusão ou discrepância na compreensão de um enunciados e contextos.

1.1.1 Léxico

> No princípio criou Deus os céus e a terra. E a terra era sem forma e vazia; e havia trevas sobre a face do abismo; e o Espírito de Deus se movia sobre a face das águas. E disse Deus**:** Haja luz; e houve luz. (BÍBLIA. *Gênesis 1*, p. 03).

Uma rápida pesquisa em dicionários[17] da Língua Portuguesa mostra que "léxico" pode ser definido como o repertório de palavras de uma língua, isto é, o vocabulário de um idioma. A constatação pode ser suficiente para o público geral, que busca informações sem grandes pretensões de análise. Entretanto, visto que esta dissertação está pautada fundamentalmente na área da Linguística, nasce a necessidade de compreender de modo mais aprofundado a significação de termos relacionados, e, por consequência, localizar com mais facilidade em quais parâmetros a temática é amparada. A discussão, portanto, é iniciada com apontamentos sobre o conceito de "palavra".

De maneira pouco esporádica, o ser humano é incentivado a acreditar no chamado "poder da palavra". Seja através de escritos antigos cultivados pelas religiões ou ainda sugestões motivacionais de profissionais que trabalham com essa metodologia, é notável que a palavra é costumeiramente colocada em posição *mágica, cabalística e sagrada* (BIDERMAN, 1998).

Tal observação pode ser ratificada com os inúmeros exemplos presentes nos textos religiosos da *Bíblia* ao tratar, por exemplo, da criação do mundo em Gênesis[18] como possibilitado a partir da palavra de Deus, ou ainda em João[19], quando é descrito como a palavra de Deus se tornou um homem, Jesus. Sob viés menos cristão, o best-seller *O Segredo*, escrito por Rhonda Byrne em 2006, apresenta concepções cujo pilar é firmado no ideal de que palavras e pensamentos positivos geram resultados assertivos. Essa proposição assume que, a partir da "lei da atração", tais comportamentos geram mudanças de vida nos âmbitos emocional, físico e financeiro.

Ademais da perspectiva mística e partindo para o lado essencialmente linguístico-comunicativo, é a palavra que permite que os indivíduos nomeiem as realidades cotidianas. Ferdinand Saussure (1857 – 1913), linguista suíço e conhecido por muitos como um dos "pais" da Linguística, sustenta a ideia de arbitrariedade da língua por acreditar que não exista uma relação natural – motivação -, entre significante e significado[20]. Sendo a língua arbitrária ou não, o fato é que, justamente, é a partir das palavras e, portanto, da língua, que o ser humano se mostra capaz de denominar o mundo e seus constituintes e, por consequência, expressar suas ideias e anseios.

---

[17] Por tratar, nesta dissertação, de assuntos que mantém diálogo aberto entre Linguística e Informática, busca-se mesclar o uso de dicionários conceituados em suas versões físicas e virtuais. Cf. referências.
[18] Gênesis é o primeiro livro da *Bíblia* e traz enunciações sobre a criação do mundo a partir de perspectiva religiosa.
[19] O texto bíblico contém Evangelhos, isto é, textos que contam a vida de Jesus Cristo. João foi um dos 12 apóstolos de Jesus e escreveu um dos Evangelhos contidos no Novo Testamento da *Bíblia*.
[20] Os dois níveis de compreensão do termo Signo linguístico - unidade fundamental de entendimento de um código em nível material (significante) e abstrato (significado), o conceito (SAUSSURE, 1969).

Canalizando a discussão para a dimensão linguístico-estrutural, cada idioma, através de suas particularidades, entende "palavra" de forma diferente. Pode-se dizer, de maneira simplificada, que palavra é um conjunto de letras e sons, ou ainda que na língua escrita é aquela unidade situada entre dois brancos, ou quem sabe, de forma um pouco mais gramatical, localizar a palavra entre o morfema – unidade mínima gramatical significativa – e o sintagma – unidade sintática que forma orações. A grande questão é que essas são assertivas da Língua Portuguesa, e uma rápida apreciação sobre outros idiomas como o Latim, o Alemão, o Russo e muitos outros que são estruturados de forma distinta, abalaria as certezas então comentadas.

A tentativa de definição de "palavra" atravessaria, portanto, não só questões estruturais como também aquelas comunicativas e, evidentemente, as representações mentais de sua significação. Múltiplas atribuições tornam a tarefa de conceituação como uma das mais complexas. Visto que o que se propõe aqui é muito mais específico e cientificamente localizado dentro dos Estudos Lexicais, adota-se, nesta dissertação, que "léxico é o conjunto abstrato das unidades léxicas da língua" (BIDERMAN, 1999).

Seguindo, então, sob os postulados de Pottier (1978) e os já mencionados de Biderman, é válido ressaltar e conceituar termos relevantes na discussão. Ao conceber que léxico é um conjunto de unidades léxicas de uma língua, o entendimento sobre o que são unidades léxicas depende sobremaneira dos autores cuja discussão tem como fundamento. Aqui, compreende-se que unidades lexicais são lexemas, lexias e lemas, sendo lexema a unidade básica do léxico de caráter abstrato, lexia a concretização dos lexemas, e lema a representação canônica dos lexemas nos dicionários. Exemplificando, as lexias *casa*, *casar* e *casamento* são lemas quando apresentadas como entradas nos dicionários, e, simultaneamente, representações concretas de unidades mínimas dotadas de significados lexicais - lexemas.

De modo mais aprofundado, Pottier (1978) e Biderman (1999) postulam que as lexias podem ser classificadas em três tipos: as simples, as compostas e as complexas. As lexias simples são aquelas formadas por uma sequência gráfica separada por dois brancos, como é o caso de, por exemplo: *guarda*, *água* e *cesta*. As compostas são as formadas por várias unidades simples ou derivadas, ligadas por aglutinação ou justaposição, e ligadas ou não por hífen como é o caso de *aguardente* e *guarda-roupa*. Por fim, as lexias complexas são aquelas também formadas por várias unidades, porém separadas por brancos, sem ligação por hífen, e que por uso constante na língua acabam por adquirir lexicalização semântica, como, por exemplo: *cesta básica*.

Especificamente, o grupo das lexias complexas apresenta restrições na seleção dos constituintes relacionadas intrinsecamente com a significação comentada. Isto é, cesta básica só tem o sentido conhecido pela sociedade - conjunto de produtos alimentícios, essenciais às necessidades nutricionais do ser humano, utilizados por uma família durante um mês – porque as lexias *cesta* e *básica* estão juntas e em posição já cristalizada pela comunidade linguística. Portanto, se a lexia complexa sofresse supressão, adição ou alteração de ordem ou tipo (**cesta*; **cesta muito básica*; **básica cesta*; **balaio básico*), o significado não seria o mesmo comentado anteriormente.

Essas combinações restritas chamadas de lexias complexas ou unidades complexas por Pottier (1974) e Biderman (1999), são para Zuluaga Ospina (1980), Corpas Pastor (1996) e Ruiz Gurillo (1997), por exemplo, unidades fraseológicas. A relação entre lexias complexas e o termo "fraseologismo" se dá justamente pela estrutura formal dessas unidades, isto é, a união dos vários corpos lexicais capazes de formar uma lexia complexa se assemelha, por analogia, ao conceito gramatical de frase – construção formada por uma ou mais palavras que contém sentido completo. Independente da nomenclatura utilizada, essas unidades são estudadas pela Fraseologia, disciplina que faz parte da Linguística e será melhor detalhada no tópico que segue.

1.2 OS ESTUDOS FRASEOLÓGICOS

Suíço por nascimento e responsável, em união ao linguista Charles-Albert Sechehaye, por organizar e editar a obra póstuma *Curso de Linguística Geral* de Ferdinand Saussure, Charles Bally (1865 - 1947) é considerado uma das figuras mais importantes da história da Linguística. Discípulo fiel de Saussure, Bally desenvolve o pensamento de seu professor através de três estudos: *Précis de Stylistique*; *Traité de stylistique française* e; *Linguistique générale et linguistique française*. É a partir de então que o linguista é considerado como precursor no estudo da Estilística como um ramo da Linguística, e que se fala pela primeira vez em *phraséologie* para tratar de alguns fenômenos sintáticos e semânticos que até então não tinham espaço diferenciado dentro da disciplina.

Observando tais fenômenos, Bally promove subsídios para uma formação teórica fraseológica, e seus preceitos rompem fronteiras encontrando forças nos estudos linguísticos da então União Soviética. Com nomes significativos como Viktor Vladimirovich Vinogradov e Nikolaj M. Šanskij, a Linguística soviética passou a assumir que as especificidades e

particularidades fraseológicas eram suficientes para fomentar uma disciplina independente com valor equivalente ao da Lexicologia[21] (KLARE, 1986).

Consequentemente as investigações fraseológicas formam toda uma escola soviética de Fraseologia a partir de 1956, e despertam o interesse de pesquisadores principalmente alemães, como Jürg Häusermann, e centro-americanos como a cubana Antonia María Tristá Pérez, por exemplo. Dentre inúmeros pesquisadores, muitos se destacam sobremaneira e, então, os estudos fraseológicos se desdobram e alcançam muitos outros países como a Espanha e a França. O espanhol Julio Casares, na década de cinquenta, é o pioneiro no seu país. Seguindo os passos do europeu, o colombiano Alberto Zuluaga Ospina em 1980, então morador da Alemanha, publica sua tese de doutorado sobre o assunto, material que acaba por se tornar um manual de Fraseologia espanhola.

Com um maior número de estudiosos discutindo a teoria, novas propostas para sua lapidação surgem de maneira muito natural. Enquanto Casares (1950) trata das unidades fraseológicas como sendo locuções e modismos, e deixando os provérbios e refrãos para a Paremiologia[22], Zuluaga (1980) amplia o preceito e considera como unidades fraseológicas as locuções e enunciados, incluindo, então, os provérbios. Dezessete anos depois, a espanhola Gloria Corpas Pastor (1997) propõe a distinção entre enunciados fraseológicos e sintagmas fraseológicos.

Assim como os autores citados, vários outros linguistas publicaram, e continuam publicando, teorias sobre a Fraseologia – como, por exemplo, Eugenio Coseriu (1977), Leonor Ruiz Gurillo (1997), Esteban Tomás Montoro del Arco (2006), Stella Esther Ortweiler Tagnin (2011) e Rosemeire Selma Monteiro-Plantin (2014), dentre outros -, e, cada um à sua maneira, usa termos específicos para lidar com as particularidades nas quais se detém. Nesse viés, mostra-se fundamental pontuar a abordagem utilizada nesta dissertação.

Entende-se, portanto, que a Fraseologia é concebida, neste trabalho, como o ramo da Linguística que se ocupa do estudo das unidades fraseológicas. Assume-se, então a conceituação de Corpas Pastor (1996) ao entender que "as unidades fraseológicas, objeto de estudo da fraseologia, são unidades léxicas formadas por mais de duas palavras gráficas em seu limite inferior, cujo limite superior se situa no nível da oração composta". Ainda de acordo com

[21] Ramo da Linguística que tem como propósito estudar cientificamente o léxico de um idioma sob diversos aspectos.
[22] Ramo da Linguística que tem como propósito estudar cientificamente as parêmias, popularmente conhecidas como provérbios, de uma língua.

os preceitos da autora, as unidades fraseológicas podem ser distribuídas em três subdivisões: Enunciados Fraseológicos (citações, parêmias e fórmulas de rotina); Colocações e; Locuções.

Para Corpas Pastor (1996), os Enunciados Fraseológicos são unidades que equivalem a enunciados completos e se subdividem em três grupos. O primeiro é o das citações, que podem ser entendidas como símbolos de dialogismo linguístico usados para amparar uma hipótese, substanciar uma ideia ou ainda exemplificar um raciocínio. O segundo das parêmias se refere ao que popularmente a sociedade conhece como provérbios e ditados populares, ou seja, frases curtas de origem popular que sintetizam um conceito a respeito de regra social ou moral – por exemplo, "De grão em grão, a galinha enche o papo"; "Deus ajuda quem cedo madruga"; "Mentira tem perna curta". Já o terceiro, das fórmulas de rotina, se encarrega das convenções de interações sociais que o falante de uma língua tem a sua disposição para utilizar em situações concretas de fala. O uso dessas fórmulas em contexto e momento adequado comumente demonstra boa educação e domínio do idioma em âmbito social. Exemplos de fórmulas são: "Bom apetite!"; "Saúde" e; "Bom dia!".

Por outro lado, as Colocações dizem respeito aos sintagmas fraseológicos que não constituem enunciados completos, isto é, precisam ser associados a outros signos linguísticos para formar atos de fala completos. Nesse sentido, pode-se afirmar que os lexemas que constituem as colocações são solidários e vêm determinados pelo uso frequente na sociedade. Os elementos dessas unidades sintagmáticas podem ser nomeados como base e colocado (WELKER, 2011), e entende-se que as bases, para gerar o sentido desejado, formam pares pré-determinados com os colocados. Dessa maneira, encontra-se na língua diversos exemplos de bases como "representar", "estar redondamente" e "mentira", que são combinados frequentemente com colocados e formam as colocações comumente reconhecidas e usadas pela comunidade: "representar um papel", "estar redondamente enganado", "mentira deslavada".

Enquanto as Colocações possuem certo grau de fixidez devido ao uso, as Locuções são também sintagmas fraseológicos que não constituem enunciados completos, mas completamente fixos no sistema (CORPAS PASTOR, 1996). Ademais, geralmente dispõem de um determinado nível de idiomaticidade, ou seja, sentido figurado. Essas Locuções idiomáticas são, no Brasil, conhecidas como expressões idiomáticas, conjunto foco deste trabalho e que será melhor detalhado no tópico seguinte.

1.2.1 Expressões idiomáticas

> As pessoas, quando corriam, antigamente, era para tirar o pai da forca e não caíam de cavalo magro. Algumas jogavam verde para colher maduro, e sabiam com quantos paus se faz

> uma canoa. O que não impedia que, nesse entrementes, esse ou aquele embarcasse em canoa furada. (ANDRADE. *Poesia completa e prosa*, p. 1993.)

Habitualmente as línguas se adequam a situações informais de fala, e especificamente os falantes da língua portuguesa usam uma quantidade volumosa de recursos linguísticos que tornam os atos comunicativos menos solenes. Tal situação é comprovada facilmente ao observar contextos de fala cotidianos cercados por gírias, jargões, interjeições, ditados, expressões e vícios de linguagem. A questão que se destaca é que as informações transmitidas a partir desses mecanismos, poderiam facilmente ser expressadas através de construções formalmente convencionais, mas a predileção dos seres humanos é inegável.

Seriam, talvez, necessários estudos psicossociais para entender a motivação por trás de tal favoritismo. Entretanto, o que não se pode negar é a presença quase religiosa dos recursos comentados. Um desses artifícios, como pode ser considerado com base no tópico anterior, é a nomeada expressão idiomática, que pode ser definida como uma lexia complexa indecomponível, de sentido conotativo e cristalizado pela sociedade linguística que a usa (XATARA, 1998).

Como detalhadamente explicitado no tópico "1.1.1 Léxico", dizer que uma lexia é complexa significa afirmar que ela é uma lexia formada por duas ou mais unidades, separadas por brancos, sem ligação de hífen, e que assume significação específica decorrente da frequência de uso. Além disso, Expressões Idiomáticas são fundamentalmente indecomponíveis, isto é, não suportam supressão ou substituição de lexias e, muito menos, alteração de ordem. Essas unidades possuem, também, sentido figurado e, portanto, sua lexicalização semântica não é denotativa. Ademais, todas essas características são parte integrante de uma expressão idiomática porque a tradição cultural, através de uso recorrente, assim determinou que seria. Em outras palavras, a unidade detém significado amplamente reconhecido pela sociedade que a manipula em atos comunicativos.

Promovendo a ilustração do que foi previamente comentado, além de trazer na epígrafe deste tópico alguns fraseologismos usados pelo poeta Carlos Drummond de Andrade, usa-se, neste momento, a expressão idiomática *pagar o pato* para exame mais detalhado dos atributos desse tipo de unidade. Analisando-a, é possível identificar que essa unidade fraseológica é composta por mais de uma lexia e separada por dois brancos – lexia complexa -, não pode sofrer alteração na ordem, substituição ou supressão das lexias (**pato pagar*, **pagar a galinha*, **quitar o pato*, **pagar*, **pato*) – indecomponível -, não significa literalmente a ação de pagar determinado valor monetário ao animal pato, mas sim o fato de ser responsabilizado por algo

injustamente – conotativa -, e tem semântica reconhecida culturalmente, sendo difícil encontrar alguém da comunidade brasileira que não entenda o significado da expressão – cristalizada-.

As Expressões Idiomáticas estão presentes nas mais diversas situações comunicativas informais do cotidiano de uma língua, porém tal assertiva não significa que uma mesma expressão estará manifestada na tradição cultural em todos os contextos sociais. Existem, por exemplo, expressões específicas de universos comunicativos que por vezes permanecem nesse ambiente, e em outros casos extrapolam os limites invisíveis e adentram outras esferas. Um bom exemplo dessa situação é o âmbito futebolístico, que possui expressões bastante específicas que permanecem somente no universo de criação e outras que avançam para os demais contextos. Algumas expressões que ilustram o cenário comentado são: *zona do agrião, comer bola, dar bola, ficar/deixar pra escanteio, pisar na bola, marcar um gol, entrar de sola, bater na trave, pendurar as chuteiras, deixar no banco, tirar o time de campo, bater um bolão,* e *vestir a camisa*.

Nessa perspectiva, é preciso destacar a não garantia de fixidez das Expressões Idiomáticas ao longo dos anos, décadas e séculos. Estudos diacrônicos comprovam facilmente que determinados fraseologismos dessa natureza podem surgir nos atos comunicativos de uma língua e permanecer por muito tempo, enquanto outros fazem aparição muito vertiginosa e logo são suprimidos. Ademais, não é esporádico, como acontece com a língua de modo geral, que as expressões sejam alteradas com o passar dos anos, e por vezes a expressão "original", ou o que a motivou, nem sequer é lembrada (LUQUE DURÁN; MAJÓN POZAS, 1998). Pode-se dizer, assim, que o que determina a permanência de uma expressão idiomática em uma língua através do tempo é a sua frequência de uso.

Independente de quanto tempo durarão, essas unidades fraseológicas surgem nas mais distintas situações sociais, mas preconizam involuntariamente as esferas linguísticas caracterizadas por caráter informal e descontraído. Um universo que compreende as características citadas é o das redes sociais. Isso acontece porque a premissa dessas plataformas é a comunicação espontânea e, em maioria, informal dos indivíduos. Sem controle de produção de textos, a língua se ramifica e, além de abrigar expressões já cristalizadas, cria diariamente outras tantas consideradas "novas"[23]. Alguns exemplos amplamente usados nesse contexto são *jantar cedo* e *passar pano*: lexias complexas, indecomponíveis, conotativas e cristalizadas no ambiente social onde são utilizadas.

[23] Nesta dissertação, são entendidas como "novas", expressões ainda não encontradas nos dicionários contemporâneos da Língua Portuguesa.

Embora considerar tais expressões como cristalizadas possa parecer uma afronta a determinados estudiosos conservadores, perceber, quantificar, e analisar essas criações mostra-se como um grande avanço no entendimento de como as tecnologias afetam ou podem impactar a Fraseologia. Negar as mudanças na língua nunca foi uma opção para a Linguística, e fechar os olhos para os novos universos comunicativos não pode ser a escolha dos Estudos Lexicais.

1.3 LINGUÍSTICA COMPUTACIONAL.

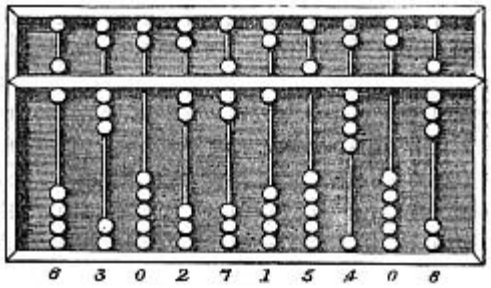
> "Sou apenas um matemático." – Alan Turing. (O JOGO DA IMITAÇÃO, 2014.)

Frequentemente promovido pela imprensa como a história do primeiro computador, o *Jogo da Imitação* (2014), com roteiro de Graham Moore e direção de Morten Tyldum, é a adaptação cinematográfica do livro biográfico *Alan Turing: The Enigma*, de Andrew Hodges. O enredo do filme narra fatos do ano de 1939, quando um discente da Universidade de Cambridge é recrutado pela recém-criada, Agência de Inteligência Britânica (MI6) para decifrar códigos criptografados nazistas. Alan e sua equipe constroem, então, uma máquina para decodificar essas informações, marcando esse momento para sempre na história mundial como um avanço tecnológico dos mais respeitáveis.

Embora tenha valor inestimável, a invenção de Alan é somente – utilizando uma expressão idiomática – *a ponta do iceberg*. O percurso histórico da invenção do computador esbarra na história da matemática (BOYER; MERZBACH, 2012) e, portanto, faz-se necessário voltar há mais de 5500 anos a.C. para falar do instrumento considerado como a primeira calculadora utilizada pelo ser humano. Com origem na palavra latina *abacus*, o ábaco foi inventado na Mesopotâmia, e posteriormente aperfeiçoado ao longo do tempo pelos povos egípcios, gregos, romanos, indianos, chineses, japoneses, americanos e russos, até chegar aos modelos educativos contemporâneos facilmente encontrados no mercado.

Salvo alterações de modelo, material e possibilidades matemáticas, o instrumento é formado por uma estrutura retangular dividida horizontalmente por uma vareta longitudinal. Além disso, como pode ser observado na figura 2, o ábaco é composto por hastes dispostas vertical e paralelamente, com um determinado número de bolas móveis em cada uma delas, tanto na parte superior da divisão horizontal (hiperbolas), quanto na inferior (hipobolas). Atualmente, o ábaco é comumente utilizado para ensinar operações de soma e subtração às crianças. Ademais, uma adaptação, nomeada como Cranmer, é frequentemente utilizada na educação básica como método de ensino do sistema numérico e aritmético para pessoas com deficiência visual.

Figura 2: Representação do ábaco em moldes ainda primitivos.



Fonte: Blog Planeta Matemática.

Muito tempo depois do primeiro registro do ábaco, Leonardo da Vinci[24] cria em 1500 a primeira calculadora mecânica que efetuava operações matemáticas simples. Pouco mais de cem anos depois, Wilhelm Schickard[25] inventa uma máquina de calcular que realizava as quatro operações matemáticas. Outros nomes como Blaise Pascal[26] em 1642, Gottfried Wilhelm von Leibniz[27] em 1671 e Charles Babbage[28] em 1839, também se destacam na história por criarem máquinas que, cada uma a sua maneira, tratavam de tentar aprimorar e desenvolver calculadoras mecânicas. Particularmente, Babbage ficou conhecido por sua máquina analítica ser considerada o primeiro computador de uso geral a utilizar somente partes mecânicas.

Relacionada intrinsecamente a esse acontecimento, Augusta Ada Byron King, Condessa de Lovelace (1815 – 1852), popularmente conhecida como Ada Lovelace, foi uma matemática e escritora de origem inglesa responsável por escrever o primeiro algoritmo a ser processado pela máquina analítica de Babbage. Graças à essas elaborações, a máquina foi capaz de computar as funções matemáticas, fazendo com que Ada seja considerada, atualmente, a primeira programadora de computadores da história.

Máquinas derivadas à esses modelos continuaram a ser desenvolvidas até a década de setenta, quando sofreram, então, substituição pelas calculadoras eletrônicas. Paralelamente à esses acontecimentos, a introdução de uma nova metodologia para as análises do censo[29] dos

[24] Cientista, matemático, engenheiro, inventor, anatomista, pintor, escultor, arquiteto, botânico, poeta e músico, o italiano Leonardo di Ser Piero da Vinci (1452 – 1519) é considerado uma das figuras mais importantes do Alto Renascimento.
[25] Inventor e professor, o alemão Wilhelm Schickard (1592 – 1635) se destacou por inventar numerosas máquinas e mapas para as áreas de astronomia, hebraico e cartografia.
[26] Matemático, escritor, físico, inventor, filósofo e teólogo, o francês Blaise Pascal (1623 – 1662) se destaca, principalmente, na geometria e na física por seus preceitos sobre Teorema de Pascal, Teoria das Probabilidades, Mecânica dos Fluídos, etc.
[27] O filósofo e polímata alemão Gottfried Wilhelm von Leibniz (1646 – 1716) é considerado um dos principais pilares da história da matemática e da filosofia. Leibniz é comumente relembrado por seus estudos sobre cálculo diferencial e integral.
[28] Charles Babbage (1791 – 1871) foi um cientista, matemático, filósofo, engenheiro mecânico e inventor inglês.
[29] Estudo estatístico referente aos dados demográficos de uma população.

Estados Unidos em 1890 influenciou sobremaneira o percurso histórico da computação (OTHERO; MENUZZI, 2005). O responsável foi o empresário norte-americano Herman Hollerith (1860 – 1929), que impulsionou o uso de cartões perfurados e máquinas leitoras desse tipo de material. Tal mecanismo era capaz de ler, automaticamente, de 50 a 220 cartões por minuto, com aproximadamente 80 dígitos cada. Em seguida, a máquina somava, multiplicava, ordenava os números e, por fim, perfurava os cartões com resultados.

O processamento de dados do citado censo de 1890 demorou três anos, o que para o padrão da época – sete anos - foi considerado um verdadeiro sucesso, colocando a máquina como aquisição certa para vários países. Quase meio século depois, em 1935 na Alemanha nazista, Konrad Zuse (1910-1995) terminou um projeto iniciado anos antes. Nomeado como Z1, o primeiro computador eletromecânico executava cálculos através de fitas perfuradas, utilizando um sistema constituído por pinos cravados em régua metálica. O modelo contou com aprimoramentos denominados como Z2, Z3 e Z4. A relação direta entre os cartões perfurados de Hollerith e a máquina de Zuse pode não ser admitida, mas a inspiração é mais do que palpável.

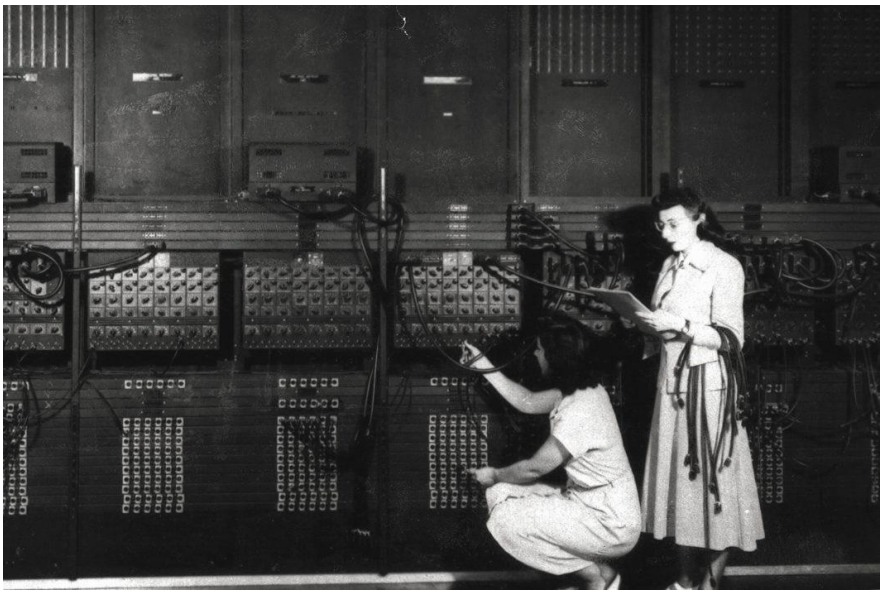
Finalmente, entre 1936 e 1939, o inglês Alan Mathison Turing (1912-1954), personagem protagonista do filme comentado no início desse tópico, desenvolve a "máquina universal". Sobre ela, Turing escreve artigos que tratavam sua produção delineando, pela primeira vez na história, a ideia de "computabilidade" ao discutir aspectos lógicos do seu funcionamento (memória, estados, transições, etc.), e entender que esse tipo de máquina seria capaz de resolver qualquer tipo de problema.

Historicamente, a competição sempre se mostrou como grande propulsora de desenvolvimento. Prova disso são os avanços tecnológicos que crescem e são aprimorados cada vez mais graças, em grande parte, à concorrência capitalista das empresas. Nesse viés, a Segunda Guerra Mundial foi uma grande impulsionadora no progresso dos computadores. Depois de Turing, os universitários norte-americanos John Vincent Atanasoff e Clifford Berry, entre 1939 e 1942, projetaram, construíram e operaram o primeiro computador eletrônico digital nomeado como *Atanasoff-Berry Computer*, e popularmente conhecido como ABC.

Por problemas com registros de patentes, o título de primeiro computador eletrônico fica com o *Electronic Numerical Integrator And Computer*, o ENIAC, em 1946, criado, a partir de pedido do exército dos Estados Unidos da América, pelos pesquisadores norte-americanos John Eckert e John Mauchly. A semelhança entre ENIAC e ABC assustava muitos pesquisadores, e em 1973, após uma batalha judicial e algumas investigações policiais, descobriu-se que

Mauchly frequentou por aproximadamente uma semana o laboratório de Atanasoff, estudou o manual da máquina ABC, e, então, construiu ENIAC com base nessas informações. Uma imagem do computador ENIAC pode ser vista na figura 3.

Figura 3: Ester Gerston e Gloria Gordon, programadoras do ENIAC.



Fonte: Reprodução/*ARL Technical Library*/*U.S. Army*.

Calcula-se que o computador ENIAC pesava cerca de 30 toneladas e ocupava aproximadamente 180m². Sua produção custou em torno de 500 mil dólares, o que representaria hoje a importância de 6 milhões de dólares, e seu sistema operacional era, embora aperfeiçoado, também baseado no sistema de perfuração de cartões. A máquina, comparada aos padrões atuais, tinha capacidade de operação de uma calculadora de mão moderna, entretanto, cumpriu o papel que lhe foi designado por 10 anos. Após sua aposentadoria, ENIAC foi desmontado e, atualmente, tem suas peças expostas pelos museus dos mais diversos países.

Quase 8 décadas depois, em 2021, esta dissertação é escrita em um *laptop* - também chamado de *notebook* -, computador compacto e leve, projetado para ser transportado e utilizado em diferentes lugares e situações. Em 1946, provavelmente poucas pessoas imaginariam os avanços tecnológicos que a sociedade presenciaria em pouquíssimo tempo, mas a realidade é que dia após dia a evolução do sistema operacional dos computadores acontece de forma rápida e, consideravelmente, imparável.

Os *upgrades* não seriam tão numerosos se tivessem permanecido efetivamente no progresso da mecânica dos computadores, e por consequência na ciência da computação[30]. A grande questão é que a "computabilidade", mencionada por Alan Turing, foi desdobrada de tal maneira que áreas até então distantes desse universo foram também beneficiadas e desenvolvidas. Uma lista completa seria praticamente impossível, visto que a quantidade de influências e apropriações é quase infinita. Entretanto, é exequível citar que a computação está presente em áreas que necessitam de soluções para processamento de dados, sejam eles escritos ou imagéticos. Nessa perspectiva, encontra-se esse tipo de tecnologia tanto na medicina, quanto na arquitetura, no ensino, e até mesmo nos esportes.

Ainda de modo relativamente tímido, a computação vem ganhando espaço também no ramo da Linguística. Compreendida como "a área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural" (VIEIRA; LIMA, 2001), a Linguística Computacional se mostra, atualmente, como um grande facilitador no que diz respeito à análise e ao tratamento do léxico.

Observando a área de modo geral, pode-se dizer que a Linguística Computacional normalmente é estudada através de duas abordagens distintas. Utilizar uma não implica no uso da outra, entretanto, não é esporádico encontrar pesquisas que aplicam a totalidade das possibilidades. Isso acontece, principalmente, em decorrência das características e dos objetivos que os estudos podem apresentar. Como esse é o caso da investigação que aqui se apresenta, este tópico sofre divisão e dois subtópicos são criados para tratar notadamente das abordagens mencionadas, sendo o primeiro nomeado como "1.3.1 Linguística baseada em *corpus*", e o segundo recebendo o título de "1.3.2 Processamento de linguagem natural".

1.3.1 Linguística baseada em *corpus*

Na introdução desta dissertação, foi explicado, de forma bastante sucinta e a fim de contextualizar o conteúdo discutido, como *corpus* linguístico - no plural *corpora* linguísticos – pode ser conceituado. Reserva-se esse momento, então, para clarificar de maneira mais específica que *corpus* pode ser compreendido como uma coletânea de textos, escritos ou transcritos de ocasiões de fala, utilizados costumeiramente em estudos sobre o comportamento linguístico da sociedade. Muitas são as fontes possíveis para criação de *corpora*, e estudiosos

[30] Ciência responsável pelo estudo das técnicas, metodologias e instrumentos computacionais, assim como de aplicações tecnológicas que informatizem os processos e desenvolvam soluções pautadas no computador.

apontam que quanto maior a extensão textual de um *corpus*, maior a possibilidade de precisão em termos de análise linguística.

Em tempos pretéritos, essas compilações eram realizadas em vias manuais. Por outro lado, com o passar do tempo e diante do desenvolvimento das facilidades informáticas, alternativas manuais tornaram-se inviáveis e, na contemporaneidade, esses arquivos são criados e mantidos de forma computadorizada. A Linguística de *corpus* é, então, a área que trata da coleta e exploração de *corpora* em arquivos de computador, oportunizando ao linguista, e outros profissionais que eventualmente trabalhem com questões de língua, novas opções para os estudos lexicais (BERBER SARDINHA, 2004).

O primeiro *corpus* linguístico eletrônico de linguagem escrita que se tem registro foi o chamado *Brown University Standard Corpus of Present-day American English*[31], popularmente conhecido como *Corpus Brown* (TAGNIN, 2018). O material, criado no ano de 1964, continha 1 milhão de palavras, o que, diante das dificuldades enfrentadas na década para informatizar coletâneas de textos, foi considerada uma quantidade apreciável. Paralelamente, os méritos de primeiro *corpus* linguístico eletrônico de linguagem falada são reservados ao professor e linguista escocês John McHardy Sinclair (1933 – 2007). Esse desenvolvimento está envolvido intrinsecamente com a criação, por parte de Sinclair e outros pesquisadores, do projeto COBUILD[32], centro de pesquisa britânico formado em 1980 pela Universidade de Birmingham em 1980 e financiado pela editora Collins.

No Brasil, a pesquisa com *corpora* eletrônicos não é recente. Biderman (1978, p. 265 – 266) discorre sobre o primeiro *corpus* brasileiro *Frequency Dictionary of Portuguese Words*, formado por 500 mil palavras do português europeu, coletadas em textos publicados entre os anos de 1920 e 1940. Ademais, a autora traz, ainda, um panorama de outros *corpora* utilizados em Estudos Lexicais no Brasil de forma precursora. Destacamos, nesse momento, aqueles compilados por quatro autores. São eles: Jean Roche (Universidade de Toulouse, França, na década de 1960); J. Hutchins (Academia Naval de Anápolis, Estados Unidos da América, anos 1970); Cléa Rameh (Universidade Stanford, Estados Unidos da América, 1972); e a própria autora Maria Teresa Biderman (Universidade de São Paulo, Brasil, 1969). Além disso, Biderman destaca uma lista de *corpora* constituídos por textos literários de autores brasileiros, analisados por uma equipe do ITA[33].

[31] Compilado por Henry Kučera e W. Nelson Francis na Universidade Brown, em Rhode Island, Estados Unidos da América.
[32] *Collins Birmingham University International Language Database.*
[33] Instituto Tecnológico de Aeronáutica.

Nessa época, os *corpora* eram usados, em grande maioria, para o ensino de línguas no que diz respeito, principalmente à criação de dicionários. Além disso, é válido ressaltar que embora fossem já eletrônicos, os primeiros *corpora* usavam sumariamente a informática. Isto é, as informações de cada *corpus* eram coletadas de forma manual e então depositadas em computador, aproveitando, assim, muito mais a capacidade de armazenamento das máquinas do que a eficiência tecnológica que elas poderiam oferecer.

Atualmente, muitos *corpora*, devido à natureza dos textos que formam as coletâneas, ainda possuem textos angariados manualmente e então transferidos para o computador. Entretanto, existe uma tendência cada vez mais forte e palpável em utilizar as facilidades informáticas presentes no cotidiano para coletar, sempre que possível, material virtual, visto que as vantagens são inúmeras e as possibilidades evoluem muito rapidamente.

Outrossim, as finalidades de uso de *corpus* mudaram sobremaneira. Enquanto no passado os *corpora* eram, em sua maioria, utilizados, como finalidade primária, o ensino de línguas e produção de dicionários, hoje eles são comumente usados para descrição linguística no amplo sentido da expressão. Nessa perspectiva, é costumeiro encontrar diariamente inúmeros novos estudos que usam como alicerce central de pesquisa a Linguística de *corpus*. Tal situação reflete eminentemente no nascimento de muitos *corpora* que buscam atender essa demanda de investigações.

Segundo Tagnin (2011), vários são os tipos de *corpora* mas essa grande gama pode ser dividida em três grupos: *corpora* de língua geral; *corpora* de língua de especialidade e; *corpora* de aprendizes. O primeiro inclui textos jornalísticos, literários e técnicos, e, por vezes, transcrições de fala. O segundo, como o nome sugere, conta com textos envolvidos em especialidades, por exemplo de culinária, de medicina, de turismo, e assim por diante. Já o terceiro envolve textos de aprendizes de língua estrangeira. Além disso, um *corpus* pode ser monolíngue, bilíngue ou até mesmo multilíngue.

Atualmente, é possível encontrar disponível na internet uma série de *corpora* prontos para uso, todos com suas particularidades de constituição e, portanto, suas características essenciais. Como exemplo, pode-se citar: Banco do Português; CetenFOLHA; CetenPUBLICO; COMPARA; *Corpus* do Português; *Corpus* Brasileiro; Lácio-Web; Linguateca; PortPopular; PHPB; TychoBrahe. De modo geral, todo *corpus* viabilizado dessa maneira conta com uma ferramenta chamada “concordanciador”, com ela é possível buscar determinada palavra escolhida para averiguar sua aparição dentro do material. Há também alguns casos nos quais tal ferramenta não é disponibilizada, ou ainda casos em que a oferta não

atende aos objetivos da investigação. Nesse caso, *softwares* externos podem ajudar imensamente no trabalho do linguista, porque, além do auxílio já mencionado, existem programas capazes de oferecer inúmeras outras ferramentas de análise de texto, como por exemplo lista de palavras, nuvem de ocorrência, gráficos estatísticos, dentre outros.

Os benefícios do uso de *corpora* linguísticos eletrônicos são inúmeros. Dentre os mais valorizados está, por exemplo, a rapidez de pesquisas relacionadas à frequência de palavras, formas ou construções da língua. Nesse viés, é possível alcançar dados decorrentes de comparações entre dois ou mais idiomas, ou ainda entre escrita e fala, e até mesmo averiguações sobre semelhanças e diferenças de uma mesma língua em diferentes períodos históricos.

No que tange à Fraseologia, a Linguística de *corpus* auxilia, além da investigação usual de frequência de termos, no entendimento de possíveis padrões de língua. Seguindo esse viés, remete-se a Sinclair (1991), que recusa o "princípio da livre escolha" por este determinar que, desde que a gramática permita, diferentes combinações de termos são possíveis, e propõe o "princípio idiomático" que entende que, em uma linguagem natural, as lexias podem apresentar certo grau de predileção por acompanhar outras específicas. Nessa perspectiva, a análise de *corpora*, a utilização de concordanciadores e, por vezes, *softwares* externos, permite tanto buscar termos específicos quanto verificar seus contextos de uso, beneficiando os estudos fraseológicos como um todo.

Ademais, destaca-se a democratização da criação de *corpus* presente na atualidade. Ao passo que antigamente a produção de *corpora* necessitava de amplos apoios financeiros e grande quantidade de tempo, vê-se que o caminho para os criar é cada vez mais viável graças à ferramentas que surgem diariamente. É exatamente nesse ponto que se abre as portas para a outra abordagem da Linguística Computacional já mencionada nesta dissertação e discutida no seguinte tópico: o processamento de linguagem natural.

1.3.2 Processamento de linguagem natural

A partir da segunda metade do século XX, começaram a ganhar espaço inúmeras pesquisas que podem ser etiquetadas como pertencentes ao universo da informática. Essas investigações incluem desde teorias para compilação de dados, desenvolvimentos de linguagens de programação, apontamentos sobre inteligência artificial, ou ainda o desdobramento de tratamentos automáticos da língua (CORI; MARANDIN, 2001). Seja de um lado ou outro, cada um a seu modo esbarra com a Linguística, porque, afinal, todas as concepções citadas atravessam a comunicação e, por consequência, a linguagem.

Como explanado de forma mais aprofundada no segundo capítulo desta dissertação, a linguagem humana é diferente daquela usada pelos computadores. Enquanto esses utilizam linguagens formais de programação, como, por exemplo, *Java*, *Python* e *PHP*, os seres humanos usam a denominada linguagem natural. Assim, cada idioma, dentro de suas especificidades, mantém uma ampla gama de características, regras de uso e estrutura, de modo que, para que um computador possa trabalhar com essa linguagem, é necessário criar mecanismos que possibilitem o entendimento durante a comunicação.

O Processamento de Linguagem Natural (PLN) – conhecido em língua inglesa como *Natural Language Processing* (NLP) e em língua francesa como *Traitement Automatique des Langues* (TAL) – pode ser entendido como uma subárea da Linguística Computacional que diz respeito, portanto, à construção de ferramentas e programas informáticos capazes de interpretar ou gerar as informações apresentadas em linguagem natural. Dessa maneira, é possível, ainda, localizar a PLN como paralelamente associada à Inteligência Artificial – IA -, que, por sua vez, estuda a capacidade e as limitações de uma máquina no entendimento da linguagem usada na comunicação humana (VIEIRA; LIMA, 2001).

Muitos são os avanços notados ao longo do tempo. Dentre eles, destaca-se, por exemplo, a tradução automática, considerada por muitos como um dos primeiros passos dados no processamento de linguagem natural. Ademais, a Linguística Computacional, no âmbito da PLN permite o desenvolvimento e construção de linguagens e ambientes de programação, ferramentas de auxílio e *softwares* adequados aos diferentes níveis de estudo da linguagem. Isto é, o tratamento automático da língua (CORI; MARANDIN, 2001) é capaz de atender, com ferramentas específicas, tanto estudos fonológicos e morfológicos, quanto sintáticos, semânticos e pragmáticos.

No que se refere aos preceitos elencados nesse capítulo teórico e na investigação aqui pretendida, entende-se que é através do processamento de linguagem natural que a Linguística de *corpus* se coloca como abordagem - e, portanto, como um estudo de língua - e não somente como metodologia (TAGNIN, 2011). Isto porque ao entender como a comunicação entre máquina e cérebro humano pode acontecer de maneira eficiente, a PLN, por meio da linguagem de programação, possibilita que novos *corpora* sejam criados e mantidos de maneira informatizada.

Mais do que facilidade mecânica, é a "computabilidade" mencionada por Alan Turing que oportuniza a autonomia que a Linguística Computacional tem como essência fundamental. A partir disso, o tratamento automático da língua abre caminhos para a formação de *corpora*

cada vez mais “computabilizados” e prontos para fornecer tanto dados premeditados quanto inesperados.

CAPÍTULO 2: ELABORAÇÃO DE UM BANCO DE DADOS POR MEIO DE COMENTÁRIOS DO *TWITTER*

Este capítulo está dedicado à exposição dos procedimentos metodológicos que orientaram o trabalho. Considerando o *Twitter* como fonte da formação do *corpus*, propõe-se, em primeiro momento, a apresentação de um percurso histórico sobre a criação da rede, e particularidades que circundam esse universo midiático-social. Ademais, são expostas as ferramentas utilizadas na constituição do banco de dados. Assim, reserva-se momento específico para explanação de fenômenos técnicos relacionados ao estudo, como preceitos sobre linguagens de programação, ambientes virtuais onde essa manipulação pode ser realizada, tópicos importantes sobre segurança cibernética, e *softwares* auxiliares na coleta e manejo de dados. Por fim, o processo de coleta e armazenamento do *corpus*, através das ferramentas escolhidas, é apresentado com o auxílio de capturas de tela.

2.1 FONTE

Inicialmente como um projeto paralelo à Odeo, empresa de *podcasts*[34] ativa entre os anos 2005 e 2017, a rede social *Twitter* surge de um compartilhamento de ideias no qual o americano Jack Dorsey propõe a criação de um serviço de troca de *status*, como, no passado, tinha-se o costume de fazer com os torpedos SMS em telefones celulares. Em um primeiro momento chamado apenas de *Status*, o protótipo do que viria a se tornar a plataforma *Twitter* tinha como função central o envio de mensagens curtas por meio do celular.

O projeto consistia em um sistema no qual a pessoa poderia enviar uma mensagem de texto para um número dizendo o que estava fazendo naquele momento, e essa seria transmitida para todos os seus contatos da lista. Nessa ideia basilar, o dono do celular perceberia o recebimento de uma atualização de *status* através de um *twich* (vibração, em língua portuguesa).

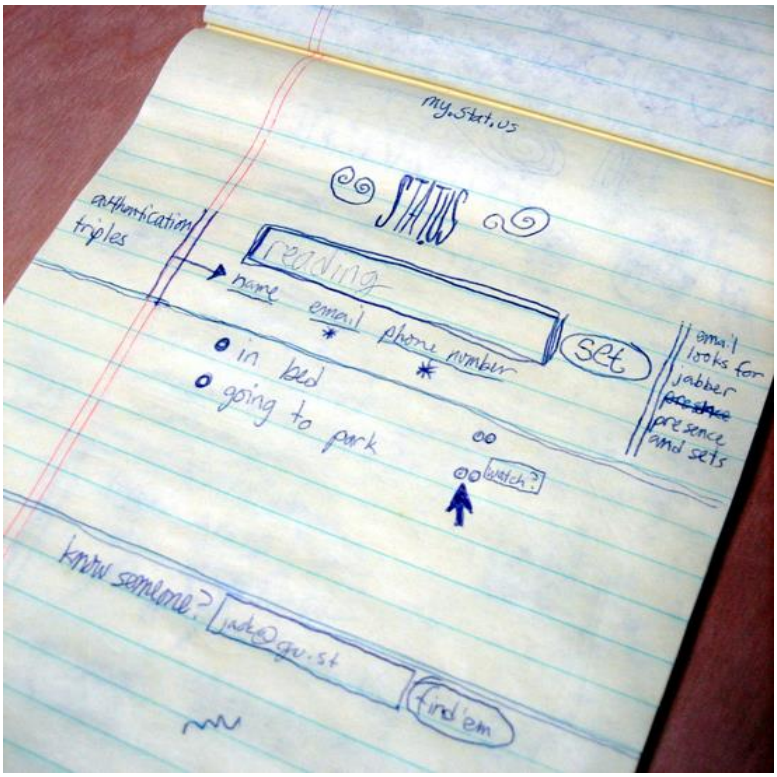
O nome inicial não agradou, e além dele cogitou-se apostar em *Sea shells*, *FriendStalker* e ainda *Jitter*. Mas o dito *twich* do aparelho celular inspirou a busca por nomes similares no dicionário, e a opção que mais chamou a atenção dos fundadores foi *Twitter*, que em inglês possui dois significados: - explosão de informações inconsequentes e; - pios de pássaros. Pensando na publicidade, foram suprimidas as vogais para que a palavra *twttr* resultasse em uma quantidade menor de teclas no celular e a ideia fosse popularizada com mais facilidade. Assim, para enviar uma mensagem de *status* via celular era preciso digitar 89887 (*twttr* nas

[34] Semelhante à um programa de rádio, *podcast* é um conteúdo de áudio produzido sob demanda e reproduzido via plataformas virtuais.

teclas). Por problemas com o registro da marca, o protótipo volta a se chamar *Twitter* e acaba por avançar em função, deixando de ser uma simples troca de SMS por celular para ser uma rede social também para computadores.

A identidade da plataforma demorou a ser definida, e, ao contrário do que muitos pensam, ela nem sempre foi conhecida como "a rede social do passarinho azul", como é popularmente chamada no Brasil. O *Twitter*, além de em princípio demorar a ter um nome definido, ao longo dos anos alterou muito a mistura de cores – carregando o verde, o branco e enfim o azul -, a estrutura da página virtual, e, claro, suas funcionalidades. Um esboço do *design* preliminarmente pensado para a plataforma pode ser conferido na figura 4. Nela pode-se observar espaços destinados a informações pessoais como "*name*", "*e-mail*" e "*phone number*", possíveis status como "*in bed*", "*going to park*", um espaço para encontrar amigos chamado de "*know someone?*", dentre outros elementos[35].

Figura 4: Caderno do CEO do *Twitter* com a primeira ideia de *design* para a rede.



Fonte: Reprodução/*Twitter*/Jack Dorsey.

No início do ano de 2006, então, o primeiro arquétipo da rede social começou a ser usado como experimento, internamente na Odeo, e graças ao seu bom funcionamento foi

35 Os elementos citados podem ser traduzidos como: *name* – nome; *e-mail* – e-mail; *phone number* – número de telefone; *in bed* – na cama; *going to park* – indo para o parque; *know someone?* – conhece alguém?

lançado oficialmente para o público em julho de 2006. Embora existam muitas polêmicas quanto às pessoas envolvidas na criação da plataforma, o que acontece também com outras redes famosas como é o caso do *Facebook*, as informações que circulam pela internet é que a origem do *Twitter* se deu através da ideia de Jack Dorsey, do financiamento de Evan Williams, do *design* de Biz Stone e a sugestão de nome de Noah Glass. Na figura 5 é possível observar o primeiro *tweet*, enviado por Jack Dorsey, ainda em forma de teste. Em tradução livre a mensagem diz: "apenas configurando meu *twttr* (*Twitter*)".

Figura 5: Primeiro *tweet* enviado por um dos fundadores da rede *Twitter*, Jack Dorsey.

Fonte: Reprodução/*Twitter*/Jack Dorsey.

O sucesso da plataforma começou, de fato, em 2007 durante o South by Southwest, festival de cinema, música e tecnologia nos Estados Unidos. A equipe do *Twitter*, naquele ano, disponibilizou duas telas no setor principal do evento, e nelas eram transmitidas mensagens trocadas pela rede social. O movimento resultou em grande êxito, e até o fim do festival foram enviados aproximadamente 60 mil *tweets* entre os participantes. No ano seguinte, a plataforma se popularizou no Brasil, e em 2011 foi lançada a sua primeira versão oficial em português. Atualmente, segundo pesquisa realizada pela empresa alemã *Statista* em 2020, o Brasil ocupa o 4º lugar no *ranking* dos países que mais acessam a rede *Twitter*.

Quanto às funcionalidades, essas foram sendo incluídas e aprimoradas com o passar do tempo e, em grande parte, aceitando sugestões e necessidades dos usuários. Uma das primeiras novidades que surgiram na plataforma, foi a possibilidade de conversar publicamente com outra pessoa específica adicionando o sinal de arroba (@) antes da identificação do usuário. Da mesma forma, surgiram as *hashtags* (#), com as quais é possível agrupar comentários sobre um mesmo assunto. Ação que facilita encontrar rapidamente informações sobre temas desejados promovendo a popularidades deles.

Tornou-se executável, também, através da função *retweet* (RT) repostar a mensagem de alguém incluindo a menção de quem fez a publicação originalmente. Ademais, como previamente comentado na introdução desta dissertação, a rede *Twitter* conta a funcionalidade *Trending Topics* (TTs). Com ela o usuário consegue acompanhar em tempo real os assuntos mais comentados ao redor do mundo, ou em uma região específica selecionada. Além disso, a plataforma tem uma série de funções menos conhecidas e/ou novas como é o caso do *Twitter List*, *Fleets*, *TweetDeck Teams*, *Thread, Topics, Qr codes*, etc.

Como pode ser observado, muito do planejamento inicial do que seria a rede *Twitter* foi alterado e aperfeiçoado. Da mesma maneira, sua função deixou de ser destinada somente ao compartilhamento de status, para englobar opiniões sobre os mais diferentes acontecimentos do mundo. Atualmente, a plataforma é famosa por compreender diariamente milhares de ponderações relacionados à política, futebol, televisão, música, cinema, relacionamento, religião, dentre outros. Mas sua fama é evidenciada ainda mais quando o que dita a discussão são os memes.

A denominação meme é amplamente utilizada virtualmente para designar uma ideia ou conceito que é difundido através da internet com velocidade elevada. A inspiração para a palavra vem do termo meme disseminado pelo escritor Richard Dawkins em seu livro The Selfish Gene, em 1976, com base nos conceitos da Memética. O autor afirma que o meme pode ser entendido como uma informação possuidora de um comportamento influenciador que multiplica e gera resultados virais. Segundo Dawkins, então, os hábitos dos seres humanos são ações naturais guiadas por genes e ações imitativas conduzidas por memes. Nas palavras do escritor:

> Os memes devem ser considerados como estruturas vivas, não apenas metafórica, mas tecnicamente. Quando você planta um meme fértil em minha mente, você literalmente parasita meu cérebro, transformando-o num veículo para a propagação do meme, exatamente como um vírus pode parasitar o mecanismo genético de uma célula hospedeira. E isto não é apenas uma maneira de falar – o meme, por exemplo, para "crença numa vida após a morte" é, de fato, realizado fisicamente, milhões de vezes, como uma estrutura nos sistemas nervosos dos homens, individualmente, por todo o mundo. (DAWKINS, 1976, p. 214).

De maneira análoga, os memes da internet, a partir do momento em que aparecem e fazem sucesso com certo número de usuários, são disseminados de maneira viral, e em pouco tempo se tornam essenciais para entender o que acontece na maioria das redes sociais. Sendo constituídos por frases, imagens, vídeos, gifs, links, sites, os memes carregam em sua maioria o humor como elemento chave, e são extremamente populares entre as gerações *millenial*[36],

[36] Também conhecida como geração Y, geração da internet, e geração milênio, geração *millenial* é um conceito da Sociologia que se refere aos indivíduos nascidos após o início da década de 1980, até, aproximadamente, o final do século.

zoomer[37] e *alpha*[38]. A partir deles, os usuários da rede *Twitter* comumente compartilham seus pontos de vista e comentários sobre o que acontece no Brasil e no mundo. É frequente, inclusive, que muitos memes nasçam na rede citada e a partir daí sejam disseminados para outras plataformas, como por exemplo *Instagram*, *Facebook* e *WhatsApp*, e ainda marquem presença em programas de televisão, jornais, revistas e publicidades, como pode ser observado na figura 6.

Figura 6: Campanha publicitária da empresa *Pizza Hut*.

Fonte: Reprodução/*Facebook*.

Nessa ocasião, a empresa *Pizza Hut* utilizou "o meme da Jéssica", como é popularmente conhecido, como recurso publicitário para incentivar seus clientes a consumirem seu produto alimentício. O citado meme surgiu de uma briga entre duas adolescentes, Lara e Jéssica, em uma escola na cidade de Alto Jequitibá, em Minas Gerais. O fato foi filmado e o vídeo amplamente compartilhado nas redes sociais. O excerto da gravação que originou o meme pode ser conferido na figura 7.

Figura 7: Captura de trecho do vídeo que originou o meme "Já acabou, Jéssica?".

Fonte: Reprodução/ *YouTube*.

[37] Também conhecida como geração Z, geração *zoomer* é um conceito da Sociologia e faz referência aos indivíduos nascidos, em média, entre a segunda metade dos anos 1990, até, aproximadamente, o início do ano 2010.
[38] Segundo a Sociologia, geração *alpha* é o grupo demográfico posterior à geração *zoomer*. Batizada com o nome da primeira letra do alfabeto grego, é a primeira a ter nascido totalmente no século XXI.

Associadas intrinsecamente aos memes, as gírias e expressões têm presença cativa nas redes sociais, e o *Twitter* é o berço de muitas delas. A figura 8 mostra o que se considera na internet como o nascimento da expressão *jantar cedo*. Na imagem é possível identificar referência a uma fala do ator Babu Santana durante sua participação no *reality show Big Brother Brasil 20*. Nesse contexto, "jantar" seria considerado um elogio, e Babu nessa ocasião foi enaltecido por grande parte do público do *Twitter* por, ao expor sua opinião, utilizar argumentos tão pertinentes a ponto de deixar o receptor sem resposta.

Figura 8: Comentário considerado como um dos primeiros a adicionar novo significado a palavra "jantar".

Fonte: Reprodução/*Twitter*.

De igual maneira, outras gírias e expressões tomam conta diariamente da rede *Twitter*. Seja *jantar cedo*; *dar biscoito*; *pprt*; *cancelar*; *passar pano*; *sco pra tu manaa*; *bomboclaat*; *old que*; *vtzeiro*; *biscoiteiro*; *o auge*; *lol; sextar*; *fic*; *quer o mundo? Eu te dou*; *deus me livre mas quem me dera*; *fingir demência*; *obrigada pelos mimos*; ou tantas outras, o fato é que a rede é palco para a atuação da língua de maneira natural e espontânea, promovendo disseminações virais e, por vezes, mutações lexicais velozes. Assim, a plataforma se mostra como fonte fértil para constituição de *corpus* que valoriza as características comentadas.

2.2 CONSTITUIÇÃO DO BANCO DE DADOS

Como previamente comentado no capítulo 1, nesta dissertação trabalha-se com pressupostos não só da Fraseologia e da Linguística de *Corpus*, mas também com os mecanismos provenientes da Linguística Computacional. Nesse viés, mostra-se válido ressaltar que programas e *softwares* específicos de coleta e análise textual têm sido frequentemente

utilizados por pesquisadores que precisam explorar uma quantidade expressiva de material textual. O uso desses mecanismos faz parte de um conjunto de novas técnicas que servem para manipular e apresentar grandes volumes de dados e obter melhores possibilidades de análise.

Assim, propõe-se, aqui neste tópico, apresentar a partir de quais ferramentas computacionais constituiu-se a base de dados desta pesquisa. Nessa perspectiva, a exposição é organizada em quatro subtópicos, sendo o primeiro aquele destinado aos apontamentos sobre a linguagem de programação *Python*, o segundo reservado para comentários acerca do ambiente de programação *Jupyter*, o terceiro para observações sobre a ferramenta de coleta de dados de mídias sociais *Socialreaper*, e o quarto para explicações sobre a chave identificadora de usuários e projetos *API Key*.

2.2.1 *Python*

Com a criação e o desenvolvimento dos computadores, nasce também a necessidade de aperfeiçoamento da comunicação entre o ser humano e esses conjuntos de componentes eletrônicos. Embora sejam dispositivos equipados com ferramentas capazes de processar dados em apenas alguns segundos, os computadores não possuem a capacidade de analisar situações novas e desconhecidas reagindo a isso de maneira espontânea. Toda e qualquer movimentação que o computador venha a realizar é fruto de uma instrução enviada primordialmente por um cérebro humano. Nessa perspectiva, a interação estrutural entre homem e máquina, que possibilita o funcionamento dessa última, é chamada de linguagem de programação.

Servindo, então, como um meio de comunicação fundamental, a linguagem de programação é um método padronizado constituído por códigos-fonte[39] que obedecem a uma série de regras sintáticas e semânticas. De forma geral, através desses códigos produzidos sob regras pré-determinadas, a linguagem de programação permite que um programador - o ser humano que mantém comunicação com a máquina - especifique de forma bastante precisa em quais dados deseja que o computador atue, de que maneira pretende que o faça e em qual ocasião opere desse modo.

Existem muitos tipos de linguagem de programação, entretanto todas elas podem ser divididas em dois grupos: aquele que se refere às linguagens de baixo nível, e o que abrange as de alto nível. Em termos gerais, os computadores só conseguem interpretar códigos quando esses estão em base binária, isto é, descritos em sistema de 0 e 1[40]. Dizer que uma linguagem de programação é de baixo nível significa afirmar que ela é produzida em base binária e,

[39] Conjunto de símbolos dispostos de forma ordenada que significam algum tipo de instrução para o computador.
[40] Exemplo: 10110000 01100001.

portanto, sofre interpretação direta pelo computador. Tal situação oferece um resultado rápido, todavia, traduzir comandos para sistemas de 0 de 1 pode ser um processo lento para os seres humanos. Em contrapartida, as linguagens de alto nível são aquelas representadas, ao invés do sistema binário, por palavras de ordem – geralmente em língua inglesa - como *print*, *copy*, *move*, etc. Essa característica auxilia o ser humano na memorização dos comandos e agiliza essa etapa da programação. Contudo, como o computador não consegue de imediato interpretar essas informações, depois de produzidos, os códigos são traduzidos para a base binária através de um compilador[41]. A escolha por um ou outro grupo vai ser definida de acordo com a preferência e as especificidades de cada programador, mas, na generalidade, as linguagens de nível alto são preconizadas visto que a tradução feita por um compilador – por ter a velocidade de uma máquina – leva menos tempo e é menos trabalhosa do que a tradução feita por um cérebro humano.

Assembly, *C++*, *Java*, *C#*, *Delphi (Pascal)*, *PHP*, *Python* e *Visual Basic* são exemplos de linguagens de programação que foram surgindo ao longo do tempo a fim de aprimorar a produtividade dos utilizadores dessas ferramentas. Procurando atender demandas diferentes, cada linguagem reflete as necessidades de sua época e, por consequência, das inovações tecnológicas que nascem e precisam de suporte para manter interação com os seres humanos.

Python[42] é uma linguagem de programação de computadores lançada em 1991 pelo matemático e programador holandês Guido Van Rossum. Em sua idealização a linguagem já tinha como objetivo principal algumas características que facilitariam o trabalho de qualquer programador ou entusiasta no assunto, são elas: linguagem fácil e intuitiva sem deixar de cumprir com funções importantes da programação; código aberto para que qualquer pessoa pudesse contribuir com seu desenvolvimento; adequação para tarefas diárias facilitando o processo de escrita de programas e; código tão inteligível quanto linguagens não informáticas. Cumprindo com suas ambições, *Python* foi desenvolvida e se tornou rapidamente uma linguagem de programação extremamente popular.

Ao combinar uma sintaxe concisa e clara com recursos poderosos, *Python* atualmente é utilizada em áreas que vão desde aquelas ligadas à engenharia, no processamento de dados científicos para análises ambientais por exemplo, até aquelas que se detém a criação de páginas dinâmicas para a *web*. Nessa perspectiva, destaca-se também o grande espaço que *Python* vem

[41] Programa de computador que, através de um código-fonte escrito em uma linguagem compilada, cria um programa semanticamente correspondente escrito em base binária. Algumas linguagens de alto nível possuem compiladores vinculados e automáticos, não precisando utilizar programas externos.
[42] Cf. https://www.python.org/.

galgando na área de linguagens. Numerosos são os trabalhos que precisam examinar grandes volumes de dados textuais ou ainda constituir dicionários, e optam, por exemplo, pelo auxílio da programação de computadores em linguagem *Python*. E é, justamente, no âmbito do tratamento e análise textual que a presente discussão encontra lugar, buscando na programação uma maneira de potencializar a composição do *corpus* aqui intencionado.

2.2.2 *Jupyter*

Como discutido no tópico anterior, visto que o cérebro humano e as máquinas não se expressam da mesma maneira, e essas não conseguem realizar ações sem instruções, a linguagem de programação é o que possibilita que um computador saiba o que, como, onde e quando fazer. Entretanto, no momento da programação de computadores, além de dispensar atenção especial ao tipo de linguagem, é necessário também que se tenha um ambiente virtual de desenvolvimento – conhecido na área da computação como IDE[43] - disponível e adequado para tal finalidade. Isto é, mostra-se fundamental que exista um espaço no qual um programador possa executar códigos e produzir seus resultados pretendidos.

Enquanto o que determina a escolha da linguagem de programação a ser adotada são dois parâmetros: as preferências do programador e; objetivos do projeto, o que estabelece a escolha por um ambiente são os dois parâmetros citados e mais um terceiro: o tipo da linguagem de programação. Muitos são os ambientes disponíveis para a programação, mas a fim de promover a eficiência da exposição aqui realizada, detém-se àqueles compatíveis com a linguagem designada nesta dissertação, *Python*.

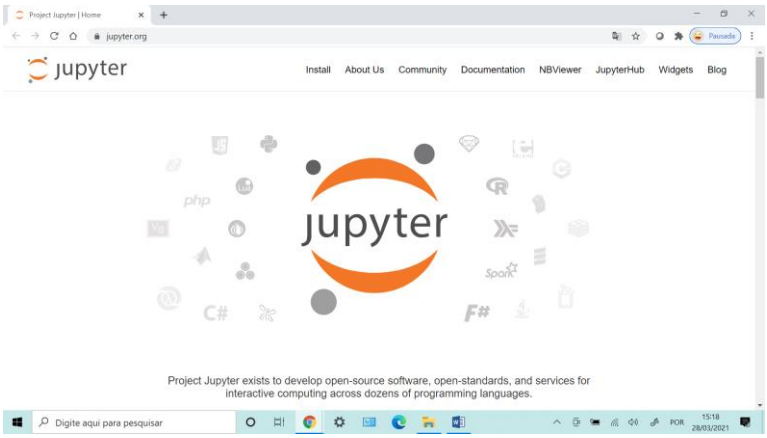
Dentre os mais conhecidos, destacam-se nesse momento cinco IDEs. *Visual Studio Code*, ou simplesmente *VSCode*, é uma IDE desenvolvida pela empresa Microsoft e oferece um espaço integralmente gratuito, inclusive para fins profissionais. Por possuir código aberto, esse ambiente é completamente personalizável e agrada muito os programadores por contar com um grande conjunto de extensões que melhoram as funcionalidades já disponíveis. Outra boa oportunidade é a *Pycharm*, pensada para funcionar com excelência tanto em Windows, quanto em Mac OS e Linux. Um diferencial dessa IDE é que o ambiente já conta com vários bancos de dados, isto é, o programador, nesse caso, não precisaria integrá-los com uma ferramenta auxiliar.

[43] *Integrated Development Environment*, em tradução livre, ambiente de desenvolvimento integrado.

Tão valorizada quanto as duas já comentadas, *Atom* é uma das IDEs mais populares devido a sua interface clara e objetiva. Esse ambiente é, como a *VSCode*, gratuito e de código aberto. Além disso conta com suporte integrado, isto é, ajuda acessível e especializada, e possui um número quase infinito de *plug-ins*[44] que melhoram a experiência e facilitam a execução de decodificações avançadas. Semelhante à *Atom*, *Spyder* é uma IDE totalmente escrita em linguagem *Python* que pode ser usada para edição, análise, depuração, entre outras funções. Está disponível de forma gratuita, tem código aberto e *plug-ins* extras. Além disso, possui suporte de codificadores e uma extensa documentação que ajuda a solucionar erros de maneira muito veloz. Completando a lista, o momento é reservado para *Jupyter*, a IDE escolhida para a constituição do *corpus* aqui almejado.

Lançado no ano de 2015, *Jupyter*[45] é um ambiente gratuito e de código aberto muito popular entre os programadores que estão iniciando a jornada em análises de dados. Seu nome faz referência às três principais linguagens de programação que o ambiente suporta: *Julia*; *Python* e; *R*, e toda a construção de imagem da marca, como pode ser observado na figura 9, é uma homenagem ao "pai da astronomia observacional" Galileu Galilei e suas anotações sobre as Luas de Júpiter[46].

Figura 9: Captura de tela da *homepage* do site da IDE *Jupyter*.



Fonte: Arquivos da autora.

Atualmente, devido à sua dinamicidade de utilização e à grande quantidade de material disponível como referência, *Jupyter* é uma das IDEs mais utilizadas e divulgadas em cursos

[44] Programa de computador usado para adicionar funções a outros programas maiores, provendo alguma funcionalidade específica.
[45] Cf. https://jupyter.org/.
[46] Descobertos em 1610 por Galileu Galilei, as Luas de Júpiter foram os primeiros objetos localizados em órbita que não a Terra ou o Sol.

online. Além de possuir bibliotecas de dados – como Pandas e Numby - que sustentam análises dessa natureza, *Jupyter* suporta trabalhos com grandes conjuntos de dados e, inclusive, funções numéricas.

Um diferencial importante de ser destacado é a existência de plataformas de ciência de dados subordinadas às premissas da *Jupyter*. Em outras palavras, é possível trabalhar com os preceitos dessa IDE em "subambientes" diversos. Dentre os mais destacados estão: *Anaconda*; *Azure Notebooks*; *Binder*; *Colaboratory*; *JupyterLab* e; *Jupyter Portable*. Para a constituição de *corpus* aqui pretendida, toda a manipulação de dados em *Python* foi realizada em *JupyterLab* por entender que esta é uma das opções mais dinâmicas e intuitivas.

2.2.3 *Socialreaper*

Integrante indissociável do universo tecnológico, byte é um tipo de unidade de medida utilizado na área da computação para categorizar tamanho e/ou quantidade. Comumente grafado em descrições de componentes, é possível citar como exemplo de uso dessa unidade de medida, o detalhamento da capacidade de armazenamento de dispositivos eletrônicos. Em termos técnicos, de acordo com os parâmetros da *American Standard Code for Information Interchange* (ASCII) , 1 byte equivale a 8 dígitos binários (bit) . Considerando a base binária 0 e 1, e elevando essa informação a oitava potência (28), 1 byte é capaz de representar 256 caracteres (28 = 256). Essa quantidade é irrisória diante da capacidade que uma máquina tem, e desenvolve diariamente, de intercambiar informações. Surgem, então, os termos derivados: kilobyte (KB); megabyte (MB); gigabyte (GB); terabyte (TB); petabyte (PB); exabyte (EB); zettabyte (ZB) e; yottabyte (YB), sendo que 1 kilobyte equivale a 1024 bytes, 1 megabyte a 1024 kilobytes, e assim sucessivamente.

Pensando de forma prática e reduzindo a expressão técnica do discurso, tudo que um computador contém dentro de sua memória – arquivos, textos, fotos, vídeos, etc. – possui um tamanho medido em byte. Da mesma forma, todos os elementos que circulam na internet possuem uma dimensão representada por bytes, e estima-se que diariamente circulem nesse ambiente aproximadamente 2,5 quintilhões de bytes em forma de dados.

Dicionarizado como uma "informação ou conjunto de informações esclarecedoras sobre uma pessoa, um grupo ou instituição" (DADO, 2021), o dado é considerado, na atualidade, um dos produtos melhor monetizados. Díspar do que grande parte da população pode imaginar, ao navegar pela internet em horários escolhidos, frequentar sites agradáveis ao gosto do usuário, participar de redes sociais preferíveis, baixar aplicativos determinados, dentre outras ações

virtuais, constituem um padrão de comportamento virtual que vem sendo correntemente utilizado por empresas especializadas em *Big Data*.

O termo pode ser traduzido de maneira livre como “grandes dados”, e se refere à uma quantidade expressiva e não estruturada de dados que são gerados virtualmente a cada segundo. O conceito de *Big Data* é sustentado por um pilar que, como mostra a figura 10, é composto por três itens: volume; variedade e; velocidade. Isso reforça a extensão da quantidade de dados, a pluralidade de conteúdos que envolve esse material, e a agilidade com que essas informações são geradas. É válido ressaltar, ainda, que com a evolução dos estudos sobre *Big Data*, surgem candidatos a novos pilares, como, por exemplo, a veracidade dos dados e o valor, isto é, a utilidade do material. Entretanto, as perspectivas são ainda recentes e buscam análises mais aprofundadas para serem entendidas como preceitos.

Figura 10: Pilares da conceituação do termo *Big Data*.

Fonte: *Data Analytics*, *Big Data*, *Data Science* - Blog Cetax.

O trabalho com *Big Data* espelha a ascensão e a metamorfose na relação entre negócio e ciência, e é considerado, na contemporaneidade, como substancial para conexões econômicas e sociais. Proporcionando novas profissões e áreas de conhecimento – engenharia de dados, ciência de dados, administração de dados, etc. -, as ferramentas de *Big Data* tem peso nas estratégias de *marketing*, auxiliando na expansão da produtividade, na atenuação dos custos, e na tomada de decisões mais eficientes. Em suma, as empresas que optam por utilizar esse novo método buscam, através de dados que refletem o padrão de comportamento de usuários, gerar valor para seus negócios, avançando sobremaneira as fronteiras do conhecimento.

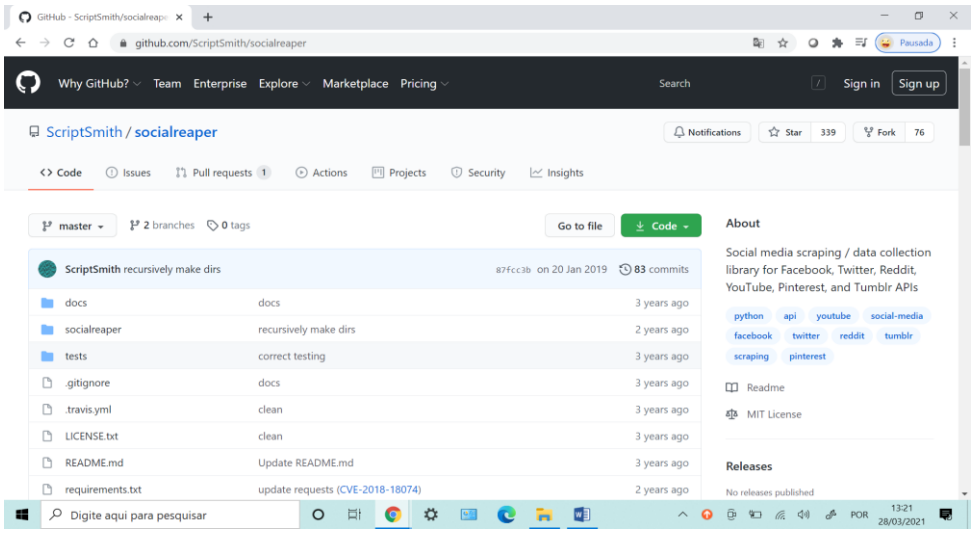
Diante do novo paradigma de transações, surgem corporações especializadas em coletar, manipular, analisar, exibir dados e, por consequência, vender esse conhecimento. Nesse viés e pensando especificamente na coleta de *Big Data*, muitos são os *softwares* capazes de viabilizar tal proposta, e a lista de canais de onde podem ser retirados esses dados inclui tanto as redes sociais, como *Facebook*, *Instagram* e *Twitter*, como também *blogs*, *sites* de notícias e todo portal que produza conteúdo gerado por um usuário, isto é, todo e qualquer site navegado por

um endereço de IP[47]. Dentre os mais conhecidos *softwares* dessa natureza, é possível citar: *Octoparse*; *Dexi.io*; *OutWit Hub*; *Scrapinghub*; *Socialreaper*; *Parsehub*; etc., e o que dita qual é mais adequado são os objetivos e aspectos da atividade pretendida.

Socialreaper[48] é um *software* internacionalmente conhecido como um *social media scraper*, em tradução livre um "raspador de mídia social", que se propõe a extrair automaticamente dados de canais de mídias sociais. Infelizmente não existem informações precisas sobre a data de criação da ferramenta. Entretanto, sabe-se que ela foi disponibilizada pela primeira vez em janeiro de 2019 através da conta virtual *ScriptSmith* do *Web Developer*[49] australiano Adam Smith[50].

A utilização do *software* é extremamente simples, e consiste em, primeiramente, fazer seu *download*. Para isso, é necessário, como mostra a figura 11, acessar a página *ScriptSmith* e, uma vez com o *software* já no computador, cria-se uma sequência de códigos em linguagem *Python* que indiquem ao programa que tipo de dados devem ser coletados e em que ambiente a ação será realizada. Em seguida, são inseridas chaves *API* - conceituação detalhada no tópico "2.2.3 *API Key*" desta dissertação – e então, o *Socialreaper* busca e apresenta em forma de lista as informações almejadas.

Figura 11: Captura de tela da conta virtual *ScriptSmith*.



Fonte: Arquivo pessoal da autora.

[47] Ao conectar um computador à uma rede de internet, ele passa a ser identificado por um endereço de IP (*Internet Protocol*). Essa identificação é representada por uma sequência de números e serve como um cadastro da máquina.
[48] Cf. https://reaper.social/ e https://github.com/ScriptSmith/socialreaper.
[49] Pode ser considerado como tipo de programador especializado em desenvolvimento de sistemas para a internet.
[50] Cf. https://github.com/ScriptSmith

2.2.4 *API Key*

Apesar de facilitar diariamente a vida de bilhões de pessoas, a internet também é um ambiente que, em muitos casos, pode se tornar perigoso. Considerado como um termo relativamente novo, crimes cibernéticos são os delitos cometidos de forma online ou principalmente online. Dentre os mais comuns, pode-se citar envio de vírus, programas e códigos maliciosos; o roubo de informações sigilosas como dados bancários, a propagação ilegal de pornografia, ou ainda condutas que reproduzam assédio, *bullying*[51], e *cyberstalking*[52].

Quando o assunto se volta para as redes sociais, a questão se torna mais delicada porque em muitos casos esbarra com o princípio básico da liberdade de expressão. Ainda que existam limites morais e legais para a disseminação de conteúdos virtuais, a livre expressão de opiniões é comumente usada como parâmetro pelos usuários das plataformas sociais quando desejam compartilhar informações sobre assuntos ou pessoas.

Pensando de forma microespacial, a Constituição da República Federativa do Brasil de 1988 impõe direitos e deveres acerca da liberdade de expressão, e leis adjacentes ratificam determinações sobre crimes cibernéticos. Entretanto, as redes sociais não estão presentes somente em solo brasileiro, e, além de precisarem cumprir determinações legais do seu território de nascimento, necessitam acatar aquelas dos países onde oficialmente disponibilizam conteúdo virtual. Portanto, cada plataforma social tem suas diretrizes internas que buscam evitar problemas legais, e, em alguns casos, as redes acabam por compartir procedimentos iguais ou semelhantes.

Embora o *Twitter*, como já mencionado, tenha conteúdo público, a coleta do material exposto nesse ambiente não possui liberação facilitada. Por questões de segurança, a rede social utiliza a chamada *Application programming interface key* (*API Key)*, que pode ser entendida como um identificador exclusivo usado para autenticar um usuário ou projeto. Isto é dizer, o usuário que deseja coletar informações do *Twitter* necessita cadastrar-se para obter uma chave *API*. Dessa maneira o sistema operacional da rede social consegue, através do cadastro, identificar usuários que estejam trabalhando com seu material a fim de evitar incidentes anônimos. Outras plataformas como *Google* ou ainda *YouTube* utilizam *API Keys*, e cada uma a sua maneira determina os critérios para cadastro e em quais casos as chaves são necessárias.

[51] Prática frequente de atos violentos e intencionais contra um indivíduo.
[52] Uso de ferramentas tecnológicas para perseguir virtualmente uma pessoa.

Tratando especificamente da rede *Twitter*, existe uma triagem bastante rigorosa para a obtenção de uma *API Key*. O usuário, que assim desejar, ingressa em um ambiente denominado *Developer Portal* destinado à desenvolvedores externos ao sistema que engloba o *Twitter*. Então, após cadastro com informações padrões pessoais, é necessário responder um extenso formulário em língua inglesa explicando e justificando o motivo pelo qual o usuário solicita uma *API Key*. Depois de ser submetido a análise, o pedido pode ser negado ou autorizado.

Sendo o pedido autorizado, o sistema gera ao solicitante uma *API Key* que consiste em três sequências de códigos. Essa chave é intransferível e precisa ser usada com extrema responsabilidade. Uma vez furtada, ela pode ser usada de forma indefinida e incontrolada, sendo o usuário cadastrado previamente o responsável por todo e qualquer problema proveniente.

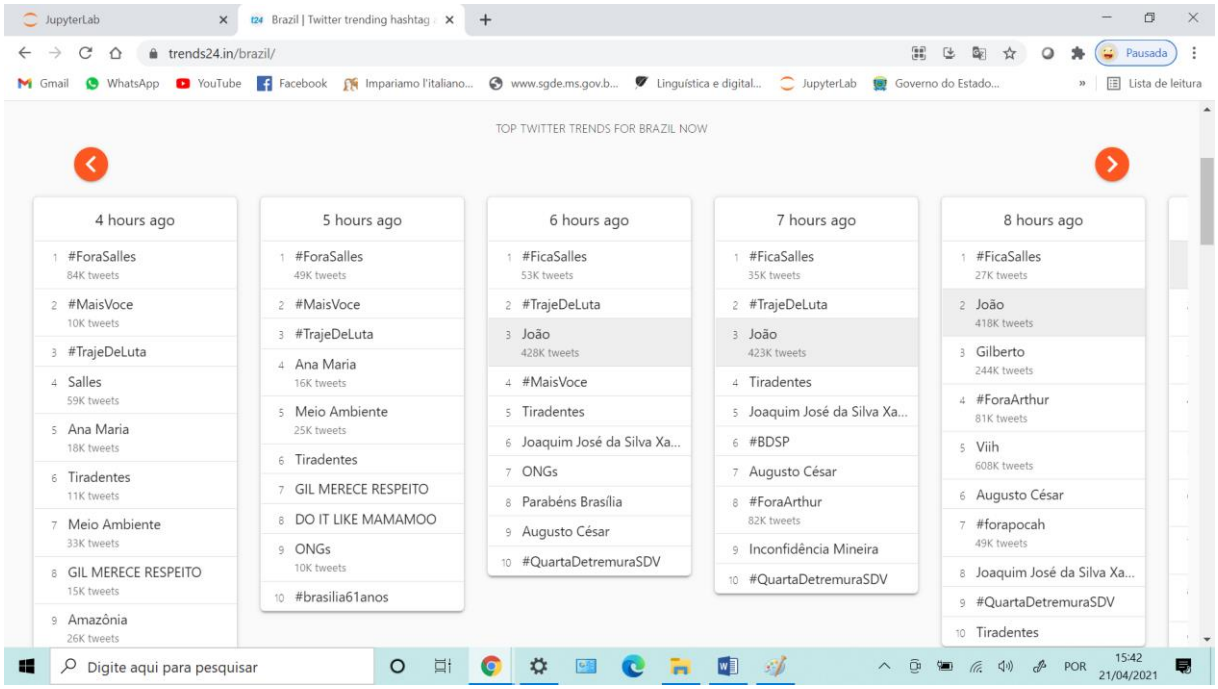
2.3 PROCEDIMENTOS

Considerando tanto as características das Expressões Idiomáticas buscadas, quanto da fonte de pesquisa – amplamente comentadas, respectivamente nos tópicos "1.2.1 Expressões idiomáticas" do "Capítulo 1: Fundamentação teórica" e "2.1 Fonte" do "Capítulo 2: Procedimentos metodológicos" -, decidiu-se realizar coletas diárias de material, a fim de constituir um *corpus* de elementos criados, pelos falantes da língua, da forma mais natural e diversificada possível. Nesse viés, o mecanismo geral da ação consistiu em armazenar eletronicamente, todos os dias, uma quantidade expressiva de comentários da rede social *Twitter*, através de um assunto mais comentado no dia em questão e, portanto, presente nos *Trending Topics* da rede social.

Visando facilitar e otimizar o trabalho, contou-se com o auxílio do website *Trends24*[53], que aponta os assuntos mais comentados das últimas 24 horas na plataforma social *Twitter*. Como pode ser observado na figura 12, esse sítio eletrônico conta com uma lista de todos os países onde a rede se apresenta, de modo que basta selecionar o país desejado e serão apresentados os *Trending Topics* diários do respectivo. Além disso, existe a possibilidade de selecionar cidades específicas dentro de cada país. Essa opção foi refutada a fim de não delimitar demasiadamente o campo de pesquisa. Assim, na seleção de localidade, escolheu-se o Brasil, pensando na rica diversidade de gírias e expressões presentes no extenso território brasileiro.

[53] Cf. https://trends24.in/brazil/

Figura 12: Captura de tela do *website Trends24.*



Fonte: Arquivo pessoal da autora.

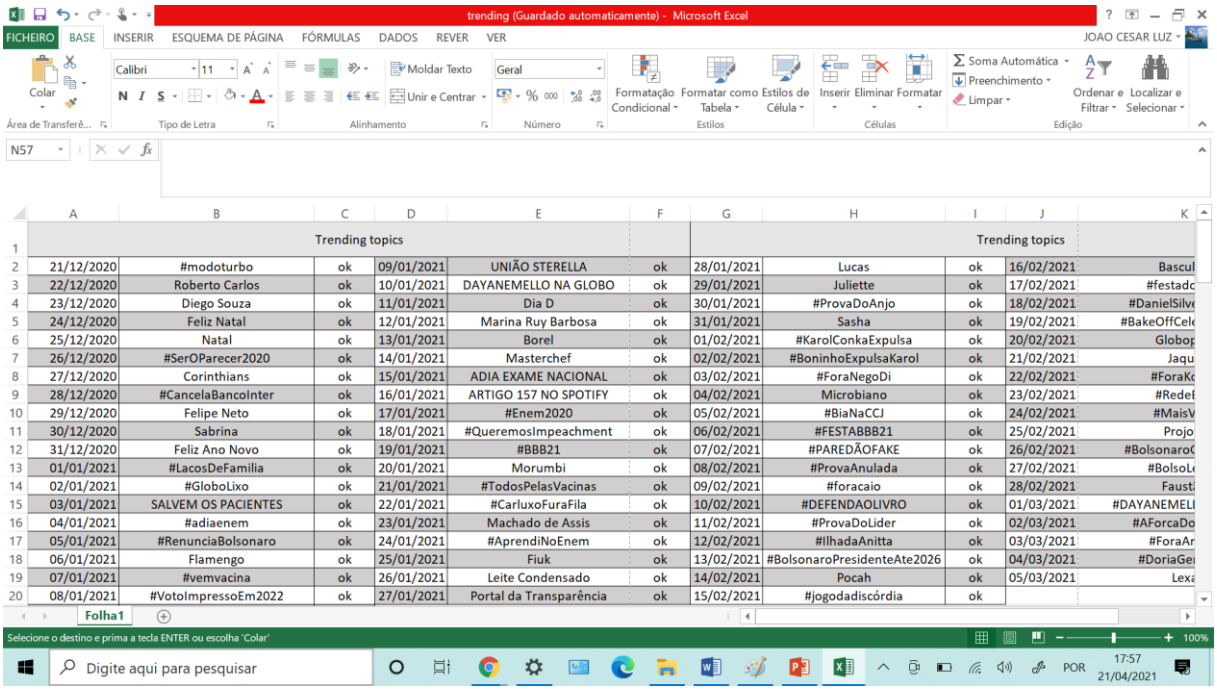
Com o propósito de organizar a coleta da maneira eficiente, elaborou-se uma tabela em arquivo *Excel* com os assuntos mais comentados escolhidos para participar da pesquisa. Em princípio, a seleção era feita a partir da quantidade de *tweets* relacionados ao tema, isto é, o assunto que tivesse maior número de comentários relacionados seria o escolhido para a pesquisa. Entretanto, com o passar do tempo, percebeu-se que tal estratégia apresentava problemáticas intrínsecas, são elas: a presença de *hashtags* estrangeiras e; o resultado em múltiplas línguas.

Mesmo trabalhando com os assuntos mais comentados no Brasil, o mundo é globalizado e muito do que acontece em outros países acaba tornando-se pauta importante em todos os continentes. Assim, é comum encontrar no *Twitter*, *hashtags* estrangeiras em posição privilegiada nos *Trending Topics* brasileiros. Além disso, outro desafio são os nomes próprios e termos mundialmente conhecidos. Para exemplificar, no dia 25 de dezembro de 2020, o assunto mais comentado foi “Jesus”, naturalmente entendível visto o que a data representa para milhões de brasileiros. Entretanto, “Jesus” está presente em diversas línguas, o que acaba por atrair comentários em múltiplos idiomas.

O mesmo acontece com “Corinthians” - nome de um time de futebol brasileiro que aparece comumente nos assuntos mais comentados em dia de jogos. Embora esteja relacionado

ao universo futebolístico, “Corinthians” em inglês se refere à um texto bíblico, o que gera resultados religiosos em língua inglesa mesclados à resultados esportivos em português. Portanto, considerando as particularidades comentadas, foram selecionados os assuntos mais comentados diários levando em conta o número de *tweets*, mas excluindo *hashtags* estrangeiras e possíveis nomes próprios e termos mundialmente utilizados. Um excerto do documento formado pode ser conferido através da figura 13.

Figura 13: Captura de tela de armazenamento em *Excel*.



Fonte: Arquivo pessoal da autora.

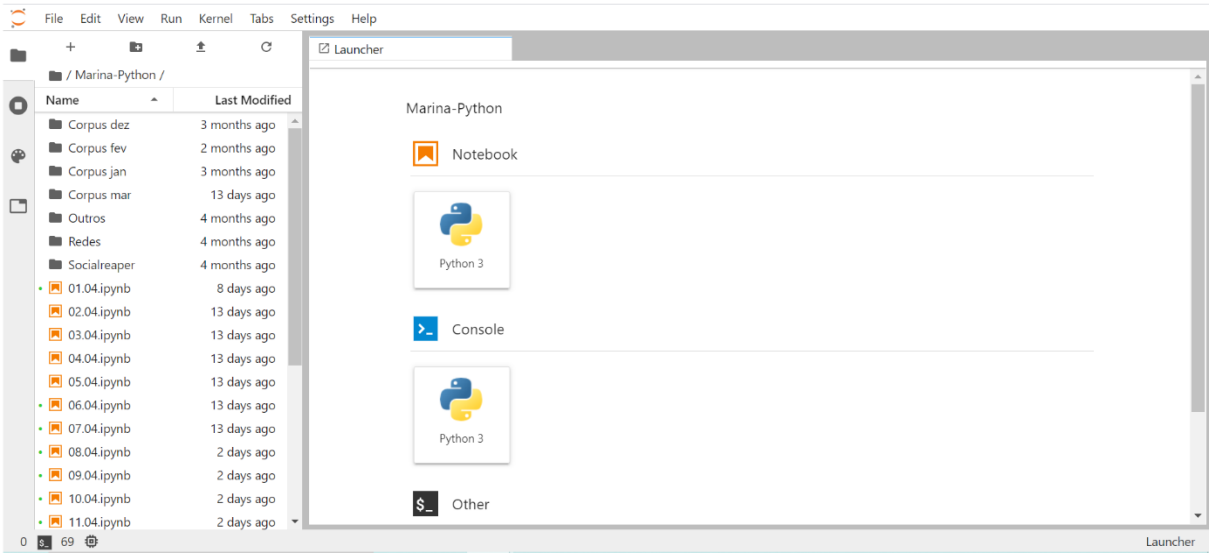
Parte-se, então, para a coleta de comentários. Como explicitado anteriormente, a linguagem de programação *Python* atualmente assiste diferentes áreas do conhecimento. Buscando atender o propósito específico da criação de *corpus*, utilizou-se nesta investigação essa linguagem em ambiente *JupyterLab*, para manejar códigos que busquem organizar a coleta de informações do *Twitter*. Nessa perspectiva, mostrou-se necessário agregar recursos externos de importação e a ferramenta escolhida foi a *Socialreaper*, que, através da “ordem” codificada e elaborada em *Python*, “raspa” os comentários pretendidos da rede social.

O ambiente de programação *JupyterLab* é uma IDE que pode ser manuseada de maneira online em qualquer navegador *web*[54] ou de modo off-line fazendo o download na área de

[54] Também conhecido como *browser*, é um programa de computador que habilita seus usuários a interagirem com documentos HTML hospedados em servidores da rede. Exemplos: *Baidu Spark Browser; Internet Explorer; Google Chrome; Microsoft Edge; Mozilla Firefox; Safari*; etc.

trabalho[55] do computador. Na pesquisa aqui apresentada, optou-se pela utilização da ferramenta de forma online. Assim, o primeiro passo para utilizá-la foi entrar no site[56] mantenedor e seguir as indicações. Em termo gerais, pode-se dizer que o *JupyterLab* se assemelha à área de trabalho padrão de um computador, embora a versão do ambiente possa trazer diferenças estéticas e funcionais. Como pode ser observado na figura 14, o lado esquerdo da área geral da ferramenta é destinado à exibição de pastas e arquivos que sofreram upload. Já o lado direito, é o ambiente de criação que conta com notebooks – aparência semelhante ao bloco de notas do *Windows*, mas com a capacidade de programação superficial -, e terminais de programação profunda[57].

Figura 14: Interface da IDE *JupyterLab.*



Fonte: Arquivo pessoal da autora.

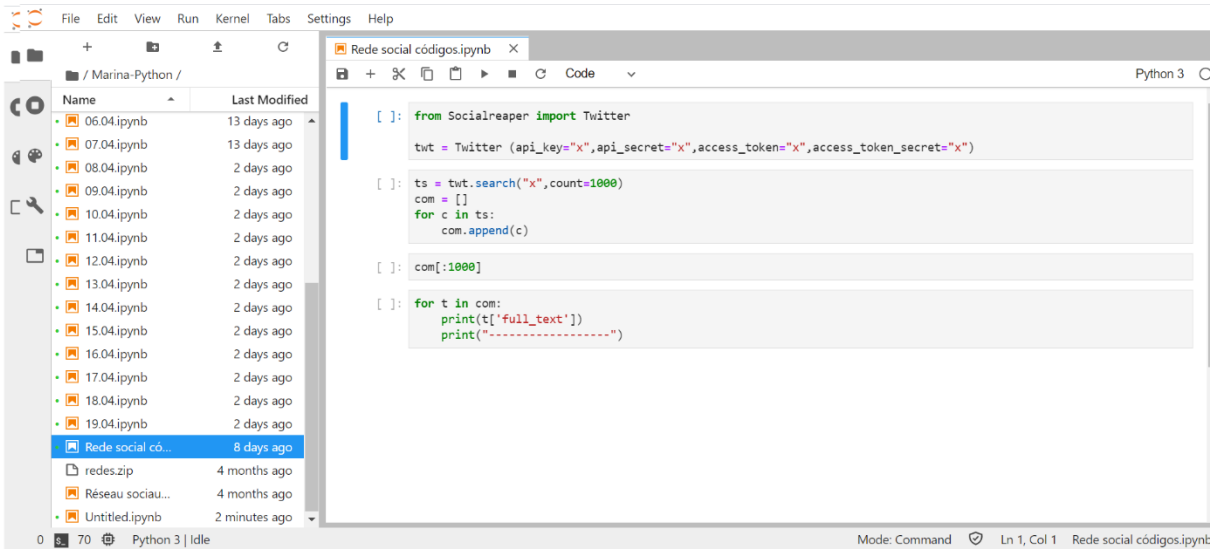
O passo seguinte foi fazer o download da ferramenta *Socialreaper* no *desktop* do computador e, na sequência, fazer seu upload no *JupyterLab*. Após ser devidamente alocado em uma pasta onde tinha-se a pretensão de realizar a programação, abriu-se um novo notebook. O *JupyterLab* quando se trata de seus notebooks, trabalha com células, ou seja, porções de informações que são efetuadas em sequência. Dessa forma, foram criados códigos que por sua vez iam sendo executados em uma cadeia de células numeradas, pela a autora, de 1 a 4.

A primeira célula informa ao computador a ferramenta a ser utilizada – "*from Socialreaper*" –, de onde ela deve importar as informações solicitadas – "*import Twitter*", e,

[55] Comumente conhecida no universo da informática como *desktop*.
[56] Cf. https://jupyter.org/.
[57] Os termos "superficial" e "profunda" foram adotados nesta dissertação para fazer referência, respectivamente, ao tipo de programação que realiza alterações externas à IDE, e ao tipo que programa comandos internos relacionados ao funcionamento base da própria IDE. Isto é, enquanto a programação superficial altera o conteúdo, a profunda é capaz de alterar a estrutura.

então, apresenta as quatro *API Keys* solicitadas pela plataforma. Na sequência, a segunda célula especifica que tipo de informações devem ser coletadas – "ts = twt.search("x") -, substituindo o "x" pelo assunto mais comentado nos *Trending Topics* do *Twitter*, e quantos comentários sobre esse tema devem ser coletados. A representação dessas ações pode ser conferida na figura 15.

Figura 15: Códigos para importação de comentários do *Twitter*.



Fonte: Arquivo pessoal da autora.

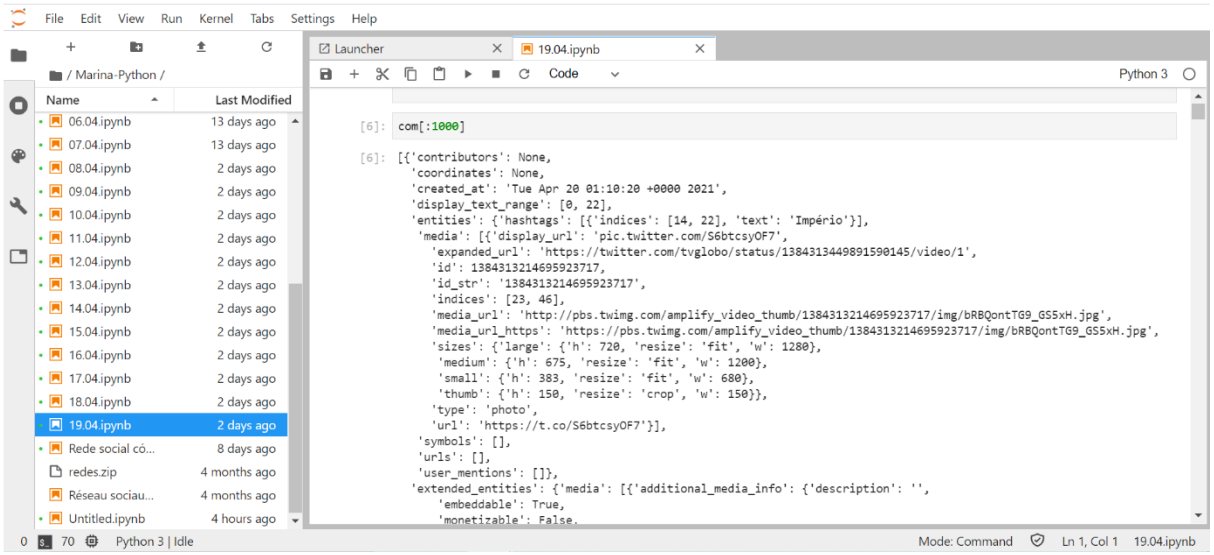
Posteriormente na terceira célula – "com[:1000]" -, pede-se efetivamente que os 1000 comentários apareçam, e aproveita-se esta oportunidade na discussão para explicar esse quantitativo. Foram testadas diferentes quantidades de comentários partindo do 100, e valores superiores a 1000 traziam como consequência a paralização tanto do ambiente de programação, quanto do *laptop* utilizado. Inúmeras experimentações foram realizadas e constatou-se que a quantidade de comentários importada é relativamente proporcional às especificações técnicas do computador que opera o processo, e a todos os componentes eletrônicos envolvidos, como a capacidade operacional da IDE, a velocidade da internet, e o fluxo de informações que navegam na rede no horário da coleta, visto que existem horários de picos.

Fazer várias coletas de 1000 comentários em um mesmo dia, totalizando um número superior total, tampouco seria viável. Isso porque a ferramenta utilizada importa os últimos 1000 comentários, e o conteúdo seria formado, assim, por repetições. Nessa perspectiva, destaca-se a importância da coleta diária não só do assunto mais comentado, como também dos comentários. Se dia 21 de abril de 2021 o assunto mais comentado é, por exemplo, "Tiradentes"

e a importação é feita somente dia 30 de abril, o resultado não é real porque serão coletados os últimos 1000 comentários do dia 30 e não do dia 21.

Retornando à terceira célula de códigos, quando se pede ao *Socialreaper* que importe 1000 comentários, a ferramenta apresenta não só os comentários, mas também uma série de informações extras. Como pode ser observado na figura 16, são trazidos dados como data, horário, nome e códigos específicos que não tem relevância para a constituição de *corpus* aqui proposta, e, de certa forma, acabam atrapalhando a visualização dos comentários.

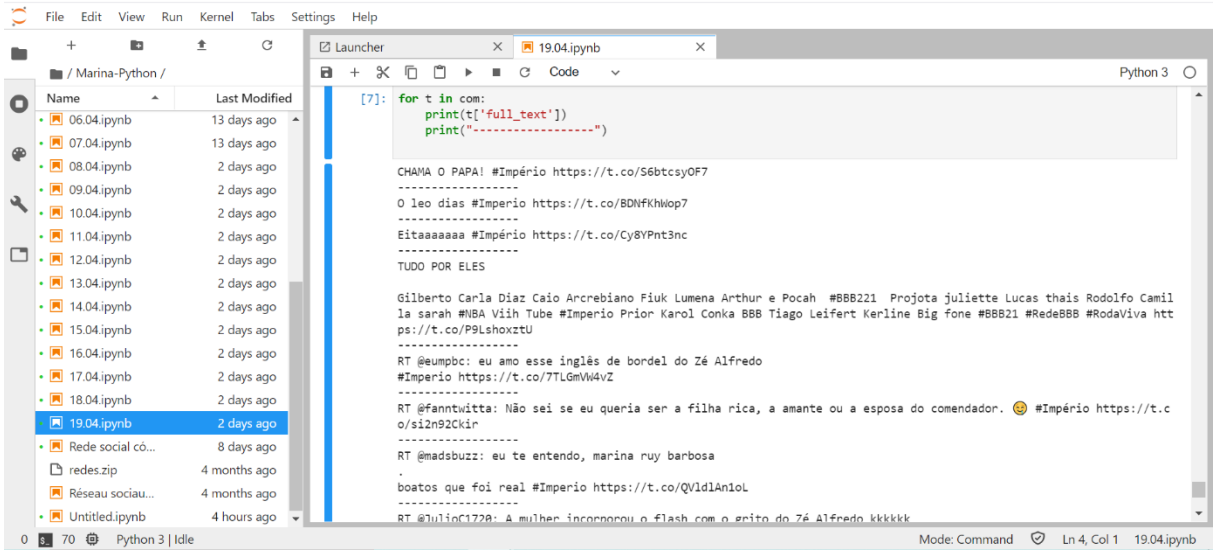
Figura 16: Captura de tela da terceira célula de códigos.



Fonte: Arquivo pessoal da autora.

Nesse viés, a quarta célula – “for t in com: print(t['full_text'] print("----------")” “limpa” as informações, como observa-se na figura 17. Com os comentários importados e livres de informações extras desnecessárias para o momento, o conteúdo textual foi transportado, através do simples comando Ctrl c + Ctrl v, à um arquivo .txt gerado pelo bloco de notas do *laptop*. A escolha pela não utilização do *Word*, ferramenta comumente usada, foi determinada pela versatilidade contida em documentos .txt. O formato permite uma série de manipulações posteriores e é o melhor aceito quando se trata de *softwares* externos.

Figura 17: Captura de tela da quarta célula de códigos.



Fonte: Arquivo pessoal da autora.

CAPÍTULO 3: ANÁLISE E DISCUSSÃO DOS RESULTADOS

Investigações lexicais que almejam estudar o comportamento linguístico da sociedade, ou de determinada parcela dela, demandam um material de pesquisa consistente que dê conta de comprovar observações e apontamentos. No caso específico deste trabalho, um *corpus* consistente se caracteriza por sua extensão e qualidade. Isto é, quanto maior a quantidade de comentários retirados da rede social *Twitter* e mais cuidadoso o tratamento desse material, mais substancial é o *corpus* e, portanto, cresce a probabilidade do estudo, como um todo, ser fidedigno, relevante e destacado para pesquisas futuras.

A tarefa de constituir *corpora* robustos para estudo lexical é, por conseguinte, um trabalho que demanda tempo. Esta pesquisa apresenta um *corpus* formado por 141.000 blocos de comentários coletados entre o período de 21 de dezembro de 2020 e 10 de maio de 2021. Preliminarmente, pode-se observar que as ferramentas de tratamento automático da língua, embora precisem considerar questões operacionais técnicas relacionadas, por exemplo, à capacidade do computador e da internet, otimizam sobremaneira o processo de formação do banco de dados. Coletar 1000 comentários diários manualmente é uma tarefa que precisaria ser desenvolvida por muito mais do que os 141 dias utilizados pela máquina.

Considerando a extensão do material e a heterogeneidade intrínseca às lexias complexas, optou-se por organizar o capítulo em diferentes tópicos. O primeiro – "3.1 Considerações sobre o uso de *Softwares* aliados no tratamento de lexias complexas" - trata sobre ferramentas de processamento de língua natural compatíveis com pesquisas sobre unidades complexas, e traz uma breve justificativa da recusa em usá-las nesta pesquisa. O segundo - "3.2 Considerações sobre as temáticas, "#" patrocinadas" e perfis robôs no *Twitter*" - apresenta discussões sobre uma série de "padrões comportamentais" nos assuntos mais comentados listados nos *Trending Topics* do *Twitter*. O terceiro, quarto e quinto - "3.3 Análise quantitativa dos dados coletados", "3.4 Análise tipológica dos dados segundo sua natureza estrutural e casos especiais" e "3.5 Análise dos dados por eixo temático", com subtópicos intitulados como "3.5.1 Eixo temático de *reality shows*", "3.5.2 Eixo temático de política" e "3.5.3 Eixo temático de novelas" – apresentam, respectivamente, dentro da delimitação escolhida para análise, investigações quantitativas, tipológicas e temáticas. Por fim, o último tópico - "3.6 Análise de possíveis candidatos a "novas" expressões idiomáticas" -, é o espaço reservado para discussões sobre expressões ainda não dicionarizadas e, portanto, material substancial para pesquisas futuras sobre possíveis relações neológicas.

3.1 CONSIDERAÇÕES SOBRE O USO DE *SOFTWARES* ALIADOS NO TRATAMENTO DE LEXIAS COMPLEXAS

Softwares específicos de análise textual têm sido muito utilizados por pesquisadores que precisam explorar uma quantidade expressiva de textos. O uso desse tipo de ferramenta faz parte de um conjunto de novas técnicas que servem para manipular e apresentar grandes volumes de dados e obter melhores possibilidades de análise. Nessa perspectiva, este tópico tem como objetivo geral organizar e apresentar informações coletadas, de diversas fontes, para facilitar e conduzir possíveis análises de dados textuais por intermédio de quatro *softwares* comumente usados no âmbito lexical. São eles: *AntConc*, *FLEx*, *Iramuteq* e *WordSmith Tools*.

AntConc[58] é um *software* gratuito desenvolvido por Laurence Anthony, professor da Universidade de Waseda, em Tóquio, no Japão. Sua última atualização, do ano de 2020, é capaz de fornecer detalhes específicos sobre um texto contido em um ou mais arquivos. Para atingir tal objetivo, a ferramenta conta com várias possibilidades de análise de texto, desde concordância, até colocados de determinado termo pesquisado. Além disso, existe a possibilidade de gerar listas de palavras de acordo com sua frequência de aparição. O *software*, conta, ainda, com a exportação dos resultados obtidos em diferentes formatos de arquivo, como, por exemplo, *HTML* e *.xlsx*.

Como pontos positivos, é possível destacar que o *AntConc* é extremamente leve – aproximadamente 4 Mb – e, por isso, dispensa instalações. Para poder utilizar o *software*, basta baixar o arquivo executável em versões para o *Windows*, *Linux* ou *Macintosh*. Ademais, ferramentas como o concordanciador e a geração da lista de palavras, são consideradas básicas e não oferecem dificuldades quanto ao uso.

De maneira geral, *AntConc* tem caráter intuitivo e o usuário pode explorar as possibilidades de forma independente até se acostumar com a interface e as alternativas. Em caso de eventuais problemáticas ou dúvidas, é possível, por meio da ferramenta de busca do *Google*, encontrar dezenas de tutoriais de vídeos e fóruns populares que auxiliem no uso. Milhares usuários, e inclusive Laurence Anthony, frequentam esse tipo de plataforma social interagindo e, por conseguinte, aprimorando a ferramenta.

Nessa linha de raciocínio, temos o *FLEx*[59] (*Fieldworks Language Explorer*) que é um *software* também gratuito, produzido pela SIL – Sociedade Internacional de Linguística[60].

[58] Cf. http://www.laurenceanthony.net/software/antconc/
[59] Cf. http://github.com/sillsdev/FieldWorks/
[60] Organização científica sem fins lucrativos, que tem como objetivo principal estudar e documentar línguas minoritárias para traduzir a Bíblia.

Dentre muitas possibilidades, essa ferramenta se destaca por ser um banco de dados projetado para auxiliar os linguistas de campo a coletar e registrar informações lexicais, fonológicas, morfossintáticas e antropoculturais, interlinearizar textos, criar e publicar dicionários (BARREIROS, 2017), contribuindo, sobremaneira, para a área lexicográfica. Ademais, *FLEx* está em constante aprimoramento, e para descarregar as atualizações do programa basta acessar o site.

De modo semelhante, o *WorldSmith Tools* é um *software* desenvolvido pelo linguista britânico Mike Scott na *University of Liverpool*, e lançado em 1996. A ferramenta conta com atualizações frequentes, e o padrão dos fornecedores é vender a versão atual, deixando as anteriores, e tecnicamente superadas, disponibilizadas de forma gratuita. Por ser um *software* que, de forma geral, apresenta possibilidades parecidas a outras ferramentas gratuitas – como a *AntConc* – se torna menos viável ao grande público, visto que o pagamento em Libra Esterlina[61] delimita os usuários.

Por sua vez, O *Iramuteq*[62] (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) foi criado na França, em 2009, por Pierre Ratinaud. É um *software* gratuito de código fonte aberto, licenciado por GNU GPL (V2), que utiliza o ambiente estatístico do *software R*[63] e, assim, como acontece com outros *softwares* de fonte aberta, pode ser alterado e expandido por meio da linguagem *Python*.

A ferramenta como *software* é considerada diferenciada das demais por proporcionar a possibilidade de criação de vários *corpora* para uma mesma análise sem prejuízo de resultados. A ferramenta francesa está apta para viabilizar diferentes análises de dados textuais tais como a lexicográfica que abrange a técnica de lematização e o cálculo de frequência de palavras. Da mesma forma, realiza análises sofisticadas e que precisam de classificação hierárquicas, inclusive com análises de similitudes.

Além disso, o *software* foi desenhado para identificar dados qualitativos, quantitativos e textuais em diferentes formas estatísticas, organizando e disponibilizando associações de palavras, e de maior relevância, e que, segundo o programa, são merecedoras de destaque. Nessa perspectiva, a ferramenta baseia-se na técnica de busca de palavras ou lematização, recurso utilizado, por exemplo, para criação de sites. Assim, o *software*, pode envolver diversas variações para uma mesma palavra nas execuções que for submetido.

---

[61] Moeda oficial do Reino Unido.
[62] Cf. https://sourceforge.net/projects/iramuteq/
[63] Cf. https://cran.r-project.org/src/base/R-3/

De modo geral, percebe-se que todos os *softwares* citados são de grande auxílio para a manipulação de dados e, embora, possam ser encontradas esporádicas incompatibilidades tecnológicas ou mercadológicas, são ferramentas de fácil instalação e uso. A breve apresentação sobre características dos programas foi em muito possibilitada pela leitura de artigos e vídeo-tutoriais referenciados. Entretanto, é ainda baixo o volume de informações sobre pesquisas que utilizem em totalidade as aplicabilidades disponíveis dos *softwares*.

De maneira específica e pensando sobre os aspectos intrínsecos às lexias complexas, foco de pesquisa no *corpus* deste trabalho, é perceptível que as ferramentas podem tanto auxiliar quanto limitar a investigação. Tratando primeiro do viés positivo, como citado no capítulo 1 desta dissertação, concordanciadores são extremamente compatíveis com expressões idiomáticas, isto porque essas unidades complexas conjuntos de itens lexicais que constituem um bloco relativamente indissociável, formado por mais de uma lexia, indecomponíveis semanticamente e, geralmente, com sentido conotativo.

Concordanciadores mostram o contexto de lexias buscadas, o que facilita a busca por expressões conotativas. Entretanto, para utilizá-los ou analisar a frequência de palavras – possibilidades comuns dos *softwares* comentados – é necessário fornecer ao programa uma lista de palavras. Desse modo, limita-se a busca. Por mais que a lista seja consistente e bastante completa, ela impedirá que o pesquisador se depare com "novas" expressões.

Portanto, utilizar *softwares* para análise de *corpus* no que tange, principalmente, às pesquisas sobre o comportamento linguístico de uma sociedade ou nicho, ao mesmo tempo que facilita a investigação por otimizar o processo, limita ao negar a possibilidade de observar novas ocorrências da língua.

3.2 CONSIDERAÇÕES SOBRE AS TEMÁTICAS, "#" PATROCINADAS" E PERFIS ROBÔS NO *TWITTER*

Como previamente comentado, primariamente foi possível perceber uma série de "padrões comportamentais" na listagem de assuntos mais comentados da rede social *Twitter*. Eles dizem respeito, principalmente, à temática dos *Trending Topics* e aos patrocínios, sejam eles oficiais ou encobertos, e o presente tópico está destinado a apresentação e discussão desses pontos.

De modo geral, pôde-se observar que diariamente os assuntos que mais apareciam na listagem foram aqueles que direta ou indiretamente estavam relacionados ao entretenimento -

incluídos o cinema, a música, os programas de televisão e as transmissões esportivas -, às datas comemorativas e à política.

Considerando os três grupos, o mais produtivo em questão de candidatos a Expressões Idiomáticas foi sem sombra de dúvidas, e como pode ser observado no tópico anterior, o de entretenimento e, especificamente, os relacionados aos programas de televisão "Big Brother Brasil", da emissora Rede Globo[64], e "A Fazenda", da emissora Rede Record[65]. *Reality shows* vem dia após dia conquistando o gosto de um número maior de brasileiros, e existem estudiosos que acreditam que tal predileção se dá pelo fato destes programas serem versões pós-modernas da encenação da vida humana (MILLAN, 2006). Tal constatação necessitaria, obviamente, de pesquisas mais aprofundadas, mas o fato é que a produtividade fraseológica dos comentários sobre *realities* mostra-se muito expressiva.

Além disso, foi possível contemplar uma quantidade significante de "#" patrocinadas, isto é, empresas que potencializam *hashtags* específicas, através de recurso financeiro. O objetivo geral é aumentar a quantidade de comentários relacionados, colocar essas informações com mais facilidade nos *Trending Topics* e, por consequência, influenciar outras pessoas a verem o assunto, se interessarem por ele, e gerarem "espontaneamente" novos comentários.

Tal mecanismo é totalmente legal e nomeado pelo *Twitter* como "*Tweets* Promovidos". De acordo com a plataforma, "os *Tweets* Promovidos são *Tweets* comuns comprados por anunciantes que desejam alcançar um grupo de usuários mais amplo ou incentivar o engajamento de seus seguidores existentes"[66]. A rede ainda destaca que quando um anunciante ou empresa paga por esse tipo de *tweet*, a situação é sinalizada com a indicação "Promovido", e em todos os outros aspectos eles contam com todas as funcionalidades de um *tweet* normal.

Infelizmente, existem organizações que promovem tal mecanismo sem contratar oficialmente o *Twitter* como anunciador. Nesse caso, empresas, sociedades, grupos ou indivíduos isolados negociam, clandestinamente, com programadores capazes de lançar milhares de comentários por minuto através de codificação informática. Essa situação ficou popularmente conhecida como "Perfil robô" ou "*bot*" – abreviação de *robot*, robô em língua inglesa - e ganhou destaque desde as últimas eleições presidenciais brasileiras de 2018. O

[64] Rede de televisão comercial aberta brasileira com transmissão nacional e internacional, fundada em 1965 pelo jornalista e empresário Roberto Marinho.
[65] Fundada em 1953 pelo empresário Paulo Machado de Carvalho, a Rede Record é uma rede de televisão comercial aberta brasileira com transmissão nacional e internacional.
[66] Cf. https://business.twitter.com/pt/help/overview/what-are-promoted-ads.html.

cenário, dentre outras coisas, influencia demasiadamente fenômenos negativos como, por exemplo, a onda de compartilhamento de *Fake News*[67].

Especificamente para este estudo, os perfis robôs influenciam na espontaneidade dos assuntos mais comentados, e, consequentemente, na naturalidade dos próprios comentários. Timidamente, começaram a aparecer no mercado plataformas específicas – como, por exemplo, o PegaBot[68] - que podem detectar *bots*, auxiliar na verificação desse material e, assim, contribuiriam para a investigação. Entretanto, novamente é ponto que necessita de ampla e cuidadosa pesquisa.

3.3 ANÁLISE QUANTITATIVA DOS DADOS

Além da elaboração de códigos bases para ferramentas digitais que auxiliassem na coleta para o levantamento de uma base de dados de textos retirados de comentários da rede social *Twitter*, esta pesquisa, visa também discutir sobre a produtividade de um *corpus* desse tipo, no que diz respeito à uma amostra da ocorrência de expressões idiomáticas, considerando a hipótese de que os usuários, no geral, se comunicam por meio de uma linguagem mais coloquial.

Dessa maneira, foi separado, de forma aleatória, o equivalente a um mês de coleta (março/2021), contendo trinta e um dias, para leitura e busca das EI. Após coletados, os textos foram armazenados no visualizador e leitor de textos "Bloco de Notas" do *software* da Microsoft Windows 10, em uma pasta compactada. Em seguida, realizou-se a leitura manual dos arquivos que apresentaram uma variação entre mil e quinhentas, e três mil linhas de texto, das quais foram levantadas um total de duzentas e trinta e uma expressões idiomáticas[69].

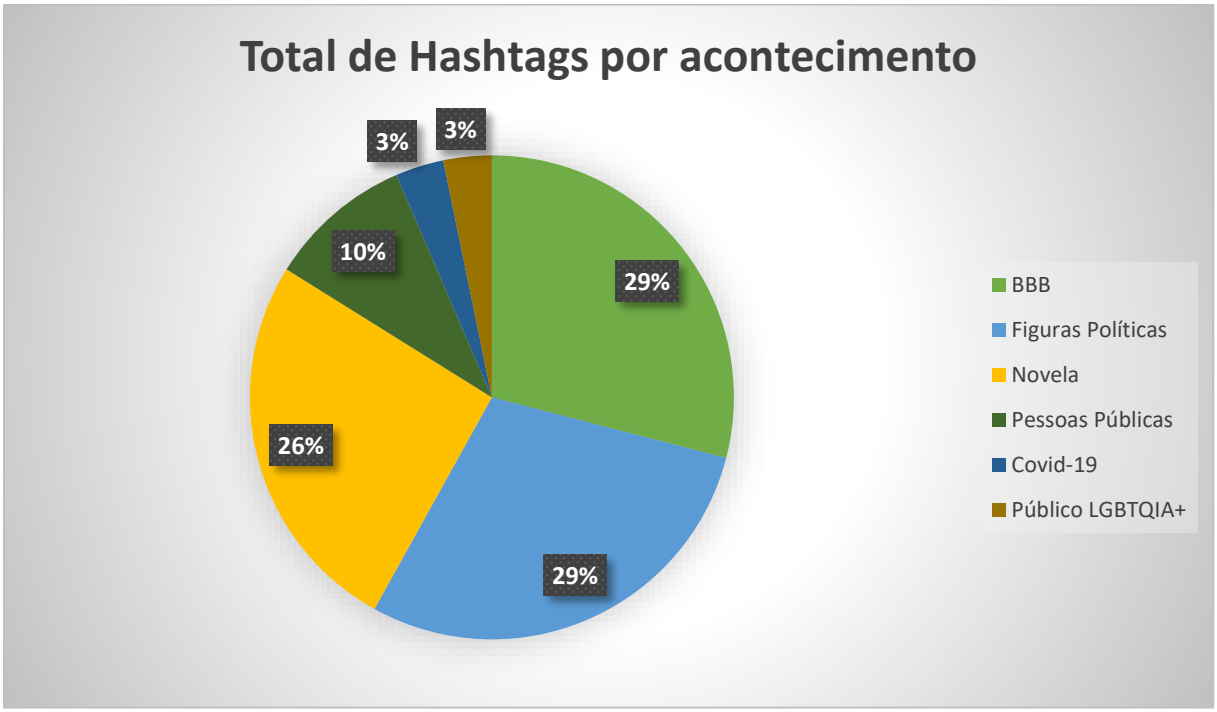
As *hashtags*[70] que aparecem no *corpus*, nesse período, foram classificadas de acordo com os temas gerais a que se referem, sendo eles: novelas, política, reality shows e outros que designam situações específicas, como: o pedido de prisão do influenciador Felipe Neto; a palavra "Lexa", que aparece tanto relacionada à apresentação da cantora de funk brasileira no reality BBB; a morte de umas personagem da série norte-americana "The 100"; o

[67] Traduzido do inglês como "notícias falsas", *Fake News* são uma forma de imprensa marrom que consiste na propagação intencional de desinformação e/ou boatos.
[68] Lançado e mantido pelo ITS Rio (Instituto de Tecnologia e Sociedade do Rio de Janeiro) e Instituto Equidade & Tecnologia. Cf. https://pegabot.com.br/.
[69] A tabela com as Expressões Idiomáticas coletadas no estudo está disponível na sessão de "Apêndices" deste trabalho.
[70] Apresentação da tabela com as *hashtags*, do mês de março, disponível na sessão de "Apêndices" deste trabalho.

posicionamento da cantora e atriz Xuxa diante dos testes de vacinas contra Covid-19[71] em presidiários; e o anúncio da primeira vacina brasileira ButanVac[72].

As esferas temáticas que se mostraram mais produtivas foram as relacionadas aos reality shows (mais especificamente o programa Big Brother Brasil e Itália) e política. De forma específica, os temas políticos giram em torno da recandidatura do atual presidente brasileiro Jair Bolsonaro, da compra de vacinas e da proposta de fechamento dos comércios, em decorrência da quarentena[73], como decorrência da Covid-19.

Gráfico 1: Total de *hashtags* por acontecimento



Fonte: Elaborado pela autora

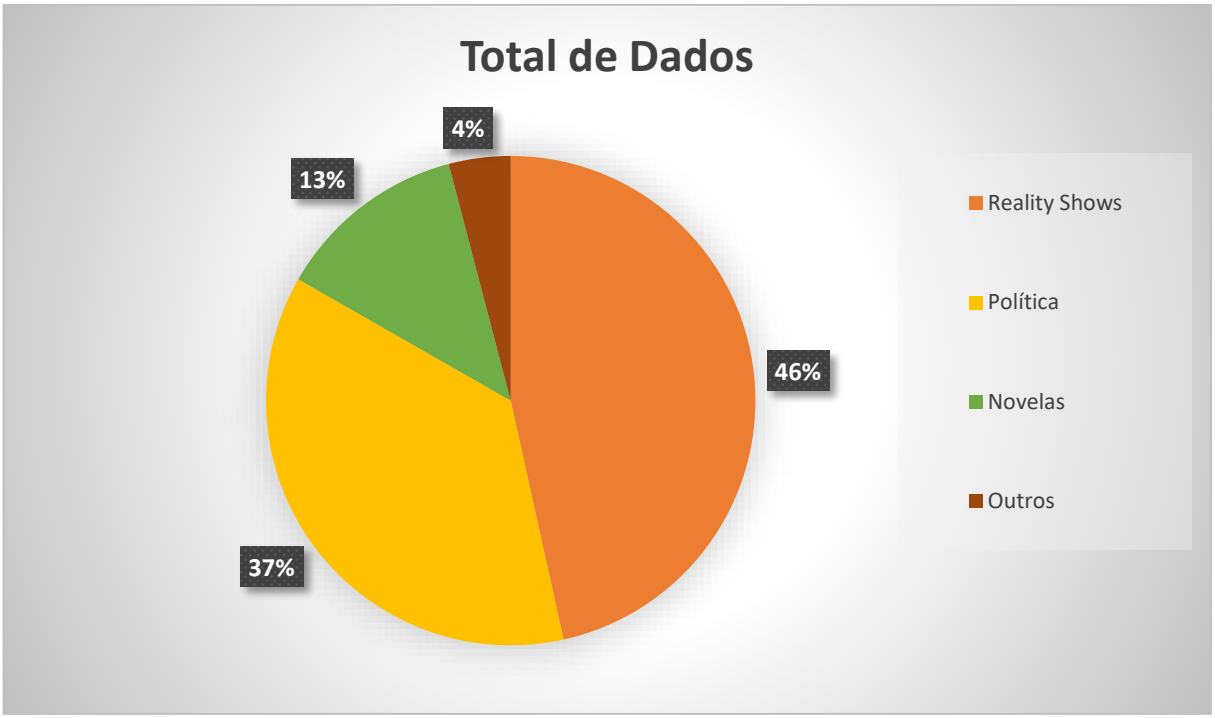
[71] A COVID-19, também nomeada como "coronavírus", é uma doença infecciosa causada pelo vírus SARS-CoV-2. No passado, variantes da doença já tinham sido detectadas em animais. Entretanto, os primeiros casos em seres humanos foram registrados em dezembro de 2019, na cidade Wuhan, província de Hubei, na República Popular da China. Rapidamente, a COVID-19 se espalhou pelo mundo, originando a chamada "Pandemia do novo coronavírus".

[72] Conhecida internacionalmente como NDV-HXP-S, a ButanVac é uma vacina contra a COVID-19. Foi desenvolvida pela Escola Icahn de Medicina no Mount Sinai, dos Estados Unidos, em parceria com a PATH – organização não governamental com foco em saúde -, e um consórcio internacional formado por instituições públicas de três países: o brasileiro Instituto Butantan, a Farmacêutica Governamental da Tailândia, e o Instituto de Vacinas e Biologia Médica do Vietnã.

[73] Medida de saúde pública que pode ser adotada durante um período de epidemia ou pandemia, e que tem como objetivo evitar a propagação de doenças infecciosas. Durante o período de quarentena é recomendado ficar o máximo de tempo dentro de casa, evitando o contato com outras pessoas e, portanto, evitando ambientes fechados e com pouca circulação de ar.

De acordo com os temas, foram coletadas cento e quatorze EI em comentários relacionados à *reality shows*, noventa à temática política, trinta e uma a assuntos sobre novelas e dez ao subgrupo "outros". É importante ressaltar que neste cálculo de levantamento não foram consideradas as repetições dessas expressões, visto que por meio da ação de *retweet* é possível que outras pessoas compartilhem o mesmo texto/*post* repetidas vezes, não caracterizando um uso autoral de fato da EI, conforme demonstra o gráfico a seguir:

Gráfico 2: Total de dados por área temática



Fonte: Elaborado pela autora.

A título de ilustração, os quadros abaixo registram as EI que foram encontradas por área temática:

Quadro 1 – Expressões Idiomáticas da área temática de "novelas".

Hashtags	**Expressões Idiomáticas**
#AForcaDoQuerer	A cobra morde o próprio rabo
#AmorDeMae	Abrir os olhos
#AForcaDoQuerer	Baixa a bola
#AmorDeMae	Cair no papinho

#AForcaDoQuerer	Cara de pau
#AForcaDoQuerer	Chegar aos pés
#AForcaDoQuerer	Comer muito arroz e feijão
#AmorDeMae	Dar o golpe do baú
#AmorDeMae	Encostar um dedo em um fio de cabelo
#AForcaDoQuerer	Ergue a cabeça
#AForcaDoQuerer	Estar aos pés
#AmorDeMae	Estar com o coração partido
#AmorDeMae	Estar mais sujo que pau de galinheiro
#AmorDeMae	Estar na cara
#AmorDeMae	Gritar aos quatro ventos
#AForcaDoQuerer	Levar chifre
#AForcaDoQuerer	Mete o pé
Thelma	Meter a boca
Thelma	Meter o pau
#AForcaDoQuerer	Morrer na minha mão
#AmorDeMae	Não ter pé nem cabeça
#AmorDeMae	Não usar o cérebro
#AForcaDoQuerer	Pegar a visão
#AForcaDoQuerer	Provando do próprio veneno
Thelma	Quebrar a cara
#AmorDeMae	Sair por cima
#AmorDeMae	Ser feita de açúcar
#AForcaDoQuerer	Só de olho
#AForcaDoQuerer	Tem que comer um quilo de fermento
Thelma	Tocar o terror
Thelma	Voar em cima
#AForcaDoQuerer	A cobra morde o próprio rabo

Fonte: Elaborada pela autora.

Quadro 2 – Expressões Idiomáticas da área temática "*reality shows*".

Hashtags	**Expressões Idiomáticas**
#DAYANEMELLOCAMPEA	Cara de Pau
#DAYANEMELLOCAMPEA	Tava na Cara
#DAYANEMELLOCAMPEA	Vamos biscoitar
#DAYANEMELLOCAMPEA	Chegar aos pés
#DAYANEMELLOCAMPEA	Ver espumando
#DAYANEMELLOCAMPEA	Dar pitaco
#DAYANEMELLOCAMPEA	Com o pé nas costas
#DAYANEMELLOCAMPEA	O choro é livre
#DAYANEMELLOCAMPEA	Brincaram com fogo
#DAYANEMELLOCAMPEA	Preparem o coração
#DAYANEMELLOCAMPEA	Ter coração explodindo
#DAYANEMELLOCAMPEA	Colocar a mão na massa
#ForaArthur	Sem conhecer a peça
#ForaArthur	Com todas as forças
#ForaArthur	Sem nem abrir a boca
#ForaArthur	Vamos em peso
#ForaArthur	Lambendo o cu
#ForaArthur	Comendo o cérebro
#ForaArthur	Faz de gato e sapato
#ForaArthur	Dando corda
#ForaArthur	Ser capacho
#ForaArthur	Nem que meu dedo tenha que cair
#ForaArthur	Foi na onda
#ForaArthur	Mala sem alça
#ForaArthur	Sem sal
#ForaArthur	Com olhos abertos

#ForaArthur	Com sangue nos olhos
#ForaArthur	Por uma luz na cabeça
#ForaArthur	Ser cachorrinho
#ForaArthur	A batata dele está fritando
#ForaArthur	Estar de saco cheio
#ForaArthur	Chorando no ombro dela
#ForaArthur	Deixar a máscara cair
#ForaArthur	Ser a vaquinha do presépio dele
#ForaArthur	Sentar o dedo
#ForaArthur	Colocar a mão no fogo
#ForaArthur	Sem cérebro
#ForaArthur	Será a última pá no enterro dele
#ForaArthur	Virar o jogo
#ForaArthur	Estar roendo as unhas
#ForaArthur	Comer os dedos
#ForaArthur	Ficar babando ovos
#ForaArthur	Cagada atrás de cagada
#ForaArthur	Cavar a própria cova
#ForaArthur	Fazer um barraco
#ForaArthur	Jogar tudo no chão
#ForaArthur	Cara de cu
#ForaArthur	Está pegando fogo
#ForaArthur	Meia boca
#ForaArthur	Lavando roupa
#ForaArthur	Esfregando na cara
#ForaArthur	A cereja do bolo
#ForaArthur	Ficar de boca fechada
#ForaArthur	Ser muito areia

#ForaArthur	Fazer a lição de casa
#ForaArthur	Refrescar a cabeça
#paredaofalso	Abrir o olho
#paredaofalso	Passar pela cabeça
#paredaofalso	A voz do povo é a voz de Deus
#paredaofalso	Com fogo nos olhos
#paredaofalso	Está caidinha por ele
#paredaofalso	Abrir o olho
#paredaofalso	Abrir o jogo
#paredaofalso	Lambendo a bunda
#paredaofalso	Lágrimas de crocodilo
#paredaofalso	Sangue nos olhos
#paredaofalso	Quebrando o barraco
#paredaofalso	É cadelinha
#paredaofalso	Ficar na mão de
#paredaofalso	Lamber pau
#paredaofalso	Rasgar o cu
#paredaofalso	Nos 4 cantos
#paredaofalso	Sangue de barata
#paredaofalso	Ombro amigo
#paredaofalso	Pelas costas
#paredaofalso	Pegar pra Cristo
#paredaofalso	Água de chuchu
#paredaofalso	Nem fede nem cheira
#paredaofalso	Bater de frente
#CarlaNoQuartoFalso	Ponto fraco
#CarlaNoQuartoFalso	Ficar alfinetando
#CarlaNoQuartoFalso	Subir a cabeça

#CarlaNoQuartoFalso	Comer o cu
#CarlaNoQuartoFalso	Correr atrás
#CarlaNoQuartoFalso	Fogo no Parquinho
#CarlaNoQuartoFalso	Apunhalar pelas costas
#CarlaNoQuartoFalso	Deixar no ar
#CarlaNoQuartoFalso	Ser planta
#CarlaNoQuartoFalso	De boca aberta
#CarlaNoQuartoFalso	Estar na boca de alguém
#BBB21Carla	Agir pelas costas
#BBB21Carla	Com o cu na mão
#BBB21Carla	Girar em torno do umbigo
#BBB21Carla	Dar a volta por cima
#BBB21Carla	Soltar a mão
#BBB21Carla	Pegar no pulo
#BBB21Carla	Cair nos braços
#BBB21Carla	Largar a mão
#BBB21Carla	Cair no papinho
#BBB21Carla	Vai lavar uma louça!
#BBB21Carla	Esperar o cabaré pegar fogo
#BBB21Carla	Estar de saco cheio
#CATBBB	Dedo duro
#CATBBB	A cara nem arde
#CATBBB	Fazer cu doce
#CATBBB	Rachar de rir
#CATBBB	Aquecer o coração
#CATBBB	Dar as caras
#CATBBB	No muro
#CATBBB	Ter sangue nas mãos

#ForaCarla	Dar gás
#ForaCarla	Lamber o rabo
#ForaCarla	Ficar babando alguém
#ForaCarla	Passar pano

Fonte: Elaborada pela autora.

Quadro 3 – Expressões Idiomáticas da área temática "política".

Hashtags	**Expressões Idiomáticas**
Lulaland	A onça vai beber água
GENOCIDA	Abaixar a cabeça
#DoriaGenocida	Abrir a boca
#EuSouOExercitoDoBrasil	Abrir mão
GENOCIDA	Abrir os olhos
#Bolsonaro2022	Aguardar as cenas dos próximos capítulos
#Bolsonaro2022	Andar de cabeça erguida
#DoriaGenocida	Apertarem o cinto
#EuSouOExercitoDoBrasil	Arrebentar a corda
#DoriaGenocida	Atingir o topo
GENOCIDA	Bater em cachorro morto
#DoriaGenocida	Botar pressão
#BolsonaroGenocidaSim	Cair na onda
Lulaland	Cara de pau
#DoriaGenocida	Chegar no ouvido
#Bolsonaro2022	Chupa essa manga
#Bolsonaro2022	Chutar cachorro morto
GENOCIDA	Colocar as coisas no eixo
#DoriaGenocida	Colocar comida na mesa

GENOCIDA	Colocar na balança
GENOCIDA	Colocar o rabo entre as pernas
#Bolsonaro2022	Conversa para boi dormir
#BolsonaroGenocidaSim	Dar bola dentro
#ConteComigoBolsonaro	Dar seu sangue
#Bolsonaro2022	Dar um tapa
#Bolsonaro2022	Dar uma mãozinha
#EuSouOExercitoDoBrasil	De braços cruzados
GENOCIDA	De meia tigela
GENOCIDA	Deixar a própria sorte
#DoriaGenocida	Deixar nas mãos
#EuSouOExercitoDoBrasil	Desenhar para todos
Lulaland	Dobrar a língua
#Bolsonaro2022	Engolir o choro
#DoriaGenocida	Entrar na sua onda
#Bolsonaro2022	Entre a cruz e a cadeirinha
#Bolsonaro2022	Entregar as ovelhas ao lobo
GENOCIDA	Esfregar na cara
#DoriaGenocida	Estamos no fogo
GENOCIDA	Estar às minguas
#DoriaGenocida	Estar cagando
#BolsonaroGenocidaSim	Estar cagando e andando
#BolsonaroGenocidaSim	Estar nas suas mãos
Lulaland	Estar no fundo do poço
#Bolsonaro2022	Enxergar o próprio umbigo
#Bolsonaro2022	Falar até papagaio fala.
#Bolsonaro2022	Farinha do mesmo saco
#ConteComigoBolsonaro	Fazer os diabos

#BolsonaroGenocidaSim	Fora da bolha
#EuSouOExercitoDoBrasil	Jogar num lamaçal
#Bolsonaro2022	Jogar sujo
#DoriaGenocida	Levar o pão para dentro de casa
#EuSouOExercitoDoBrasil	Mais claro que a água
#Bolsonaro2022	Mandar no pedaço
#DoriaGenocida	Mãos sujas de sangue
#Bolsonaro2022	Máscaras caindo
#BolsonaroGenocidaSim	Não largar o osso
GENOCIDA	Não vai colar
GENOCIDA	Não valer o que o gato enterra
#DoriaGenocida	Nas mãos
GENOCIDA	No chão
Lulaland	O golpe tá aí, cai quem quer
GENOCIDA	O pai tá on
#Bolsonaro2022	O pau vai torar
#DoriaGenocida	Pagar um preço
GENOCIDA	Passador de pano
GENOCIDA	Passar a mão
GENOCIDA	Passar a mão na cabeça
GENOCIDA	Peidar fino
#ConteComigoBolsonaro	Peidar na farofa
#DoriaGenocida	Precisa cair
GENOCIDA	Preparar o lombo
#Bolsonaro2022	Quebrar a cara
#DoriaGenocida	Rasgar as cortinas
#PLdoGasOriginal	Sentar o dedo
#ConteComigoBolsonaro	Separar os homens dos meninos

#DoriaGenocida	Ser duro
GENOCIDA	Ser uma barca furada
#DoriaGenocida	Te jogar em um poço
#BolsonaroGenocidaSim	Tem que cair
#Bolsonaro2022	Tentar derrubar
#BolsonaroGenocidaSim	Ter a batata assando
GENOCIDA	Ter as mãos sujas de sangue
#Bolsonaro2022	Ter na mão
GENOCIDA	Ter sangue nas mãos
#Bolsonaro2022	Ter um pingo
#BolsonaroGenocidaSim	Tirar as maçãs podres
#Bolsonaro2022	Tirar o cavalinho da chuva
#Bolsonaro2022	Valer a pena
#Bolsonaro2022	Virar uma barata
#Bolsonaro2022	Voltar a pastar
Lulaland	A onça vai beber água

Fonte: Elaborada pela autora.

Quadro 4 – Expressões Idiomáticas da área temática “outros”.

Hashtags	**Expressões Idiomáticas**
Lexa	Serviu muito
Lexa	É uma bolha
Lexa	Perdi o chão
Lexa	Gastar meu réu primário
#FelipeNetoNaCadeia	Sentir o gosto do próprio veneno
#FelipeNetoNaCadeia	Colocar as cartas na mesa
#FelipeNetoNaCadeia	Meter o pé
#FelipeNetoNaCadeia	Dar de cara
#FelipeNetoNaCadeia	Borrar as calças

#FelipeNetoNaCadeia	Soltar a mão
Lexa	Serviu muito

Fonte: Elaborada pela autora.

Dentre as EI coletadas, cento e vinte e seis são somáticas[74], ou seja, possuem em sua estrutura alguma lexia que remeta às partes do corpo humano e/ou animal, representando 50% dos dados. Esse número expressivo de ocorrências poderia justificar-se pelo contexto extralinguístico dos sujeitos, já que esses se valem dos fenômenos imagéticos a sua volta para construir significado, como evidenciam as pesquisas da fraseóloga Mellado Blanco, que pontua que:

> […] o homem utiliza-se do que há de mais concreto ao seu redor, do seu próprio corpo, para fazer referência a outros fenômenos mais abstratos, geralmente apresentado com um forte conteúdo expressivo, que ele associa a atitudes, gestos ou movimentos realizados por seu corpo. Deste modo, o estudo das imagens e metáforas que o homem usa para verbalizar seus sentimentos, nos permite ver a chave dos fenômenos que fazem parte do seu ambiente objetivo e que resultam subjetivamente mais relevantes. (MELLADO BLANCO, 2004, p. 31, tradução nossa)

Enquanto as outras apresentam lexias que designam animais, alimentos e objetos, como: corda e poço. Deve-se lembrar que algumas expressões pertencem a mais de uma tipologia temática, como por exemplo a EI "lágrimas de crocodilo", que pode ser tanto considerada uma zoônima (em razão do uso de um termo que designa um animal – crocodilo), quanto uma somática (em razão do fluído corporal – lágrima).

Para fins de ilustração, seguem abaixo os quadros com dados referentes às EI somáticas, zoônimas e de alimentos:

Quadro 5 - Expressões Idiomáticas Somáticas

Hashtags	**Expressão Idiomática**
#DAYANEMELLOCAMPEA	Cara de Pau
#DAYANEMELLOCAMPEA	Tava na Cara
#DAYANEMELLOCAMPEA	Chegar aos pés
#DAYANEMELLOCAMPEA	Com o pé nas costas
#DAYANEMELLOCAMPEA	O choro é livre
#DAYANEMELLOCAMPEA	Preparem o coração

[74] Nesta pesquisa, considera-se o exposto por Mellado Blanco (2004) e a conceptualização complementar proposta por Marques (2007), que inclui ao conceito de expressões somáticas, além de nomes que se referem a partes do corpo humano e animal e seus órgãos, "líquidos e fluídos corporais", como sangue e lágrimas.

#DAYANEMELLOCAMPEA	Ter coração explodindo
#DAYANEMELLOCAMPEA	Colocar a mão na massa
#AForcaDoQuerer	Morrer na minha mão
#AForcaDoQuerer	Ergue a cabeça
#AForcaDoQuerer	Mete o pé
#AForcaDoQuerer	Só de olho
#AForcaDoQuerer	Chegar aos pés
#AForcaDoQuerer	Cara de pau
#AForcaDoQuerer	Estar aos pés
#AForcaDoQuerer	A cobra morde o próprio rabo
#AForcaDoQuerer	Levar chifre
#AForcaDoQuerer	Provando do próprio veneno
#ForaArthur	Sem nem abrir a boca
#ForaArthur	Abrir os olhos
#ForaArthur	Lambendo o cu
#ForaArthur	Comendo o cérebro
#ForaArthur	Nem que meu dedo tenha que cair
#ForaArthur	Com olhos abertos
#ForaArthur	Com sangue nos olhos
#ForaArthur	Por uma luz na cabeça
#ForaArthur	Chorando no ombro dela
#ForaArthur	Sentar o dedo
#ForaArthur	Colocar a mão no fogo
#ForaArthur	Sem cérebro
#ForaArthur	Estar roendo as unhas
#ForaArthur	Comer os dedos
#ForaArthur	Cagada atrás de cagada
#ForaArthur	Cara de cu

#ForaArthur	Meia boca
#ForaArthur	Esfregando na cara
#ForaArthur	Ficar de boca fechada
#ForaArthur	Cara de pau
#ForaArthur	Dedo duro
#ForaArthur	Refrescar a cabeça
#DoriaGenocida	Abrir a boca
#DoriaGenocida	Nas mãos
#DoriaGenocida	Deixar nas mãos
#DoriaGenocida	Chegar no ouvido
#DoriaGenocida	Mãos sujas de sangue
#DoriaGenocida	Estar cagando
#DoriaGenocida	Abrir mão
#paredaofalso	Abrir o olho
#paredaofalso	Passar pela cabeça
#paredaofalso	Com fogo nos olhos
#paredaofalso	Abrir o olho
#paredaofalso	Lambendo a bunda
#paredaofalso	Lágrimas de crocodilo
#paredaofalso	Sangue nos olhos
#paredaofalso	Quebrando o barraco
#paredaofalso	Ficar na mão de
#paredaofalso	Lamber pau
#paredaofalso	Rasgar o cu
#paredaofalso	Sangue de barata
#paredaofalso	Ombro amigo
#paredaofalso	Pelas costas
GENOCIDA	Passar a mão

GENOCIDA	Ter as mãos sujas de sangue
GENOCIDA	Colocar o rabo entre as pernas
GENOCIDA	Peidar fino
GENOCIDA	Esfregar na cara
GENOCIDA	Ter sangue nas mãos
GENOCIDA	Passar a mão na cabeça
GENOCIDA	Abrir os olhos
GENOCIDA	Preparar o lombo
GENOCIDA	Abaixar a cabeça
#Bolsonaro2022	Enxergar o próprio umbigo
#Bolsonaro2022	Andar de cabeça erguida
#Bolsonaro2022	Dar uma mãozinha
#Bolsonaro2022	Ter na mão
#Bolsonaro2022	Quebrar a cara
#Bolsonaro2022	Engolir o choro
#CarlaNoQuartoFalso	Abre teu olho
#CarlaNoQuartoFalso	Abrir o olho
#CarlaNoQuartoFalso	Subir a cabeça
#CarlaNoQuartoFalso	Subir pra cabeça
#CarlaNoQuartoFalso	Comer o cu
#CarlaNoQuartoFalso	Apunhalar pelas costas
#CarlaNoQuartoFalso	De boca aberta
#CarlaNoQuartoFalso	Estar na boca de alguém
Lulaland	Dobrar a língua
Lulaland	Cara de pau
Lulaland	Estar no fundo do poço
#BBB21Carla	Agir pelas costas
#BBB21Carla	Com o cu na mão

#BBB21Carla	Girar em torno do umbigo
#BBB21Carla	Soltar a mão
#BBB21Carla	Comendo o cu
#BBB21Carla	Cair nos braços
#BBB21Carla	Largar a mão
#BBB21Carla	Cair no papinho
#BBB21Carla	Ficar com o cu na mão
#EuSouOExercitoDoBrasil	Abrir mão
#EuSouOExercitoDoBrasil	De braços cruzados
#BolsonaroGenocidaSim	Estar nas suas mãos
#BolsonaroGenocidaSim	Não largar o osso
#BolsonaroGenocidaSim	Estar cagando e andando
#CATBBB	Dedo duro
#CATBBB	A cara nem arde
#CATBBB	Fazer cu doce
#CATBBB	Aquecer o coração
#CATBBB	Dar as caras
#CATBBB	Ter sangue nas mãos
#PLdoGasOriginal	Sentar o dedo
#FelipeNetoNaCadeia	Sentir o gosto do próprio veneno
#FelipeNetoNaCadeia	Meter o pé
#FelipeNetoNaCadeia	Dar de cara
#FelipeNetoNaCadeia	Soltar a mão
#ConteComigoBolsonaro	Peidar na farofa
#ConteComigoBolsonaro	Dar seu sangue
#ForaCarla	Lamber o rabo

Fonte: Elaborada pela autora.

Quadro 6 – Expressões Idiomáticas Zoônimas

Hashtag	Expressão Idiomática
#AForcaDoQuerer	A cobra morde o próprio rabo
#ForaArthur	Faz de gato e sapato
#ForaArthur	Ser cachorrinho
#ForaArthur	Ser a vaquinha do presépio dele
#paredaofalso	Lágrimas de crocodilo
#paredaofalso	É cadelinha
#paredaofalso	Sangue de barata
GENOCIDA	Não valer o que o gato enterra
GENOCIDA	Bater em cachorro morto
#Bolsonaro2022	Tirar o cavalinho da chuva
#Bolsonaro2022	Entregar as ovelhas ao lobo
#Bolsonaro2022	Chutar cachorro morto
#Bolsonaro2022	Conversa para boi dormir
#Bolsonaro2022	Falar até papagaio fala.
Lulaland	A onça vai beber água

Fonte: Elaborada pela autora.

Quadro 7 – Expressões Idiomáticas com nomes de alimentos.

Hashtags	**Expressões Idiomáticas**
#AForcaDoQuerer	Comer muito arroz e feijão
#AForcaDoQuerer	Tem que comer um quilo de fermento
#ForaArthur	A batata dele está fritando
#ForaArthur	Ficar babando ovos
#ForaArthur	A cereja do bolo
#paredaofalso	Água de chuchu
#Bolsonaro2022	Chupa essa manga
Lulaland	A onça vai beber água
#EuSouOExercitoDoBrasil	Mais claro que a água

#BolsonaroGenocidaSim	Tirar as maçãs podres
#BolsonaroGenocidaSim	Ter a batata assando
#CATBBB	Fazer cu doce
#ConteComigoBolsonaro	Peidar na farofa
#AmorDeMae	Ser feita de açúcar

Fonte: Elaborada pela autora.

3.4 ANÁLISE TIPOLÓGICA DOS DADOS: SEGUNDO SUA NATUREZA ESTRUTURAL E CASOS ESPECIAIS

Xatara (1998, p. 17), ao dissertar acerca da tipologia das EI, ressalta que elas podem ser classificadas a partir de duas formas: a) segundo sua natureza estrutural; e b) por casos especiais. A primeira, refere-se à *"natureza morfossintática que confirma o princípio de complexidade lexical"*. Enquanto que o segundo tipo, refere-se às expressões de língua francesa que podem ser adaptadas nos mesmos contextos da língua portuguesa contemporânea. Em sua dissertação de mestrado, Euzébio (2021), apresenta dois quadros que resumem tais tipologias apresentadas pela autora, como pode-se ver a seguir:

Quadro 8: Tipologia das EI segundo sua natureza estrutural

1) sintagmas nominais: referem-se às expressões que, em uma oração, exercem a função de substantivo – "*cintura de pilão", "amigo da onça"*[75].
2) sintagmas adjetivais: referem-se às expressões que exercem a função de adjetivo e podem conter construções paralelas ou não – "*são e salvo", "altos e baixos".*
3) sintagmas adverbiais: referem-se às expressões que exercem a função de advérbio – "*por baixo do pano", "em cima do muro"*.
4) sintagmas verbais: são expressões que correspondem a um verbo. Podem ser formadas por: • V + SN – "*colocar o coração", "entregar o ouro*". • V + ADJ + SN – "*ter a última palavra*".

[75] Os exemplos utilizados pela autora nos dois quadros foram retirados do artigo de Xatara (1998).

• V + preposição + SN – "*dar no pé", "escapar pelos dedos*".
5) A autora atenta-se também para o caso das **EI elípticas** "nas quais não se explícita um dos elementos do sintagma frasal: *[...] estar nas alturas*". • Os sintagmas frasais, de maneira geral, são exclamativos e formados por uma oração – "*é o fim da picada!"* ou por frases nominais – "*e daí, eu com isso", "pra cima de mim?"*.

Fonte: Retirado de Euzébio (2021, p. 34) com base em Xatara (1998)

Diante do *corpus* textual elaborado neste trabalho em conjunto com a teoria tipológica segundo a sua natureza estrutural, proposta por Xatara (1998), as Expressões Idiomáticas coletadas enquadram-se nas seguintes classificações:

A) sintagmas nominais: referem-se às expressões que exercem a função de adjetivo no nível oracional. Exemplos:

- (1) "Carla ir pro quarto falso e ver o povo falando mal dela e voltar com **sangue nos olhos."** (Expressão com função adjetiva, significando: "irada").
- (2) "Paredão **água de chuchu**!!!" (Expressão com função adjetiva, significando: "indiferente", "sem graça").

B) sintagmas verbais: são expressões que correspondem a um verbo. Podem ser formadas por:

V + SN – JOGAR + SUJO

- (3) "Estão **jogando sujo** para tentar derrubar Bolsonaro." (Significando: trapacear para conseguir algo).

V + preposição + SN – AGIR + pelas + COSTAS

- (4) "É fácil se dizer arrependido depois de tudo que falou e como **agiu pelas costas** após o retorno da Carla". (Significando: fazer algo para alguém sem o conhecimento dele).

C) EI elípticas: "nas quais não se explícita um dos elementos do sintagma frasal." Exemplo:

- (5) "Vcs da gazeta do povo são a imprensa ruim. Que bate em cachorro morto, e **passa a mão** em genocida." (Há supressão da lexia 'cabeça' – passar a mão na cabeça, significando "não cobrar as responsabilidades").

- (6) “Camila e João **no muro**.” (Há supressão de parte da expressão – “estar em cima do muro”, significando indecisão).

Quadro 9: Tipologia dos casos especiais das EI segundo Xatara (1998)

1) EI alusiva: aquelas que necessitam da incursão de um conhecimento enciclopédico para serem entendidas e os fatos sejam esclarecidos para decodificar a expressão: *“só acredito vendo”* (relacionado a São Tomé, que era incrédulo).
2) EI análoga: refere-se àquela que apresenta alguma semelhança, porém possui sentido distinto: *“pôr em dia”* (relacionado a fofocar) e *“pôr em xeque”* (desafiar). (RIVA; CAMACHO, 2002, p. 209).
3) EI apreciativa: tem efeito de sentido pejorativo: *“farinha do mesmo saco”, “filhinho de papai”.*
4) EI comparativa: apresenta em sua estrutura elementos comparativos: *“pegajoso como carrapato”, “liso como sabonete”.*
5) EI deformada: expressão que apresenta trocadilhos: *“onde o Judas perdeu as botas”, “ver-se em papos de aranha”*, versão erudita da palavra palpos
6) EI hiperbólica: expressa exagero: *“matar cachorro a grito”.*
7) EI irônica: expressa ironia por meio da contrariedade: *“rápido como uma tartaruga”.*
8) EI negativa: são utilizadas na negativa: *“não abrir mão”, “não esquentar a cabeça”.*
9) EI numérica: possui, na sua composição léxica, um numeral: *“matar dois coelhos com uma cajadada só”.*
10) EI situacional: é empregada em uma situação social precisa ou desencadeada por uma situação específica. Geralmente, designam uma provocação: *“nem mais um pio”, depois você me conta”.*

Fonte: Retirado de Euzébio (2021, p. 35), com base em Xatara (1998).

Considerando os casos especiais pontuados por Xatara (1998), nota-se no *corpus* de pesquisa, as seguintes ocorrências:

A) alusivas: aquelas que necessitam da incursão de um conhecimento enciclopédico para serem entendidas. Exemplos:

- (1) "**estaremos** em 2022 outra vez polarizados, **entre a cruz e a caldeirinha...**" (Segundo o Dicionário de Expressões Correntes, de Orlando Neves, edição da Editorial Notícias, refere-se àquelas pessoas que estavam para morrer ou até mesmo mortas, na qual, em épocas antigas eram colocados um crucifixo sobre a cabeça dos doentes e aos seus pés uma caldeira de água benta. Com o tempo, atenuou-se o sentido da expressão e passou a significar: estar em um dilema ou ter que escolher entre duas opções igualmente ruins).
- (2) "Cris querendo ser mãe para dar o **golpe do baú**". (Expressão utilizada desde o século XVIII, pois nessa época as pessoas costumavam guardar suas joias, roupas e artigos de valor em baús. É utilizada para caracterizar pessoas que se casam ou engravidam por interesse, para que possam receber as heranças de seus parceiros)[76].
- (3) "Vamos derrubar essa farsa que é esse presidente **de meia tigela**". (Expressão originada da monarquia portuguesa, pois os funcionários da coroa recebiam seus alimentos de acordo com o trabalho que realizavam. Ou seja, pessoas consideradas "menos importantes" recebiam apenas meia tigela de comida. É utilizada com significado de insignificância, algo sem qualidade)[77].

B) análogas: refere-se àquela que apresenta alguma semelhança em sua estrutura, porém possui sentido distinto

- (4.1) "Vai trabalhar Senador e **abra mão** dos seus benefícios em prol dos necessitados". (Significado – desistir, renunciar).
- (4.2) "Ele pode **abrir o jogo** pra Carla". (Significado – revelar, contar ou mostrar algo que está escondido).
- (4.3) "Carla tem que ir pro quarto secreto pra **abrir o olho**". (Significado – perceber, ficar atenta).
- (5.1) "que horas qui é pra **colocar a mão na massa**???". (Significado – iniciar uma atividade).

[76] Referência histórica para compreender o sentido da expressão "golpe do baú": Dias Costa, Eva (2 de março de 2019). "A posição sucessória do cônjuge sobrevivo no direito português: a propósito da lei 48/2018, de 14 de agosto".
[77] Cf. em <https://www.significados.com.br/meia-tigela/>. Acesso em: 03 de janeiro de 2022.

- (5.2) “Projota, Caio e Rodolffo **colocando a mão no fogo** pelo Arthur, sem ter ouvido a conversa entre Arthur e Juliette...”. (Significado – confiar muito em alguém).
- (6.1) “Acho que é bom **colocar as cartas na mesa**, falar umas verdades.... as vezes é legal. É necessário”. (Significado – esclarecer as coisas).
- (6.2) “só Lula voltar ao cenário, começou **colocar as coisas no eixo**”. (Significado – alinhar o que está fora de ordem).

C) apreciativa: tem efeito de sentido pejorativo. Exemplo:

- (7) “Loulaa e FHC é tudo **farinha do mesmo saco**, o PSDB é um lixo comunista! ” (Expressão teve sua primeira aparição na língua escrita em latim “Homines sunt ejusdem farinae”, com base metafórica na referência histórica do ensacamento de farinha, no qual, as de boa qualidade ser posta em sacos separados, para não serem confundidas com as de qualidade inferior, utilizada para significar que os iguais andam juntos, os bons com os bons e os maus com os maus).

D) hiperbólica: expressa exagero. Exemplos:

- (8) “É sério que tem um governador **rasgando as cortinas** do palácio Bandeirantes”. (Significa – desesperado, sem saber como agir).
- (9) “Gente, vamo acordar, de onde vocês tiraram que a Carla vai voltar **quebrando o barraco**?” (Significa – fazer confusão).

F) negativa: são utilizadas na negativa. Exemplos:

- (10) “Eles msm estão morrendo. Mas **não largam o o$$o!**”. (Significa – renunciar algo que aprecia).
- (11) “Essa história do flagrante do Sandro foi uma das coisas mais ABSURDAS que eu já vi na vida. **Não tem pé nem cabeça**”. (Significa – algo que não faz sentido).
- (12) “Firula e Sandro sendo presos por **não usarem o cérebro**”. (Significa – não utilizar de inteligência).
- (13) “Agora que o Lula apareceu, aí peida fino de medo. **Não vai colar**”. (Significa – atitude não levada a sério ou não tida como verdadeira).
- (14) “Quero o impeachment do genocida mesmo sabendo que Mourão ou Lira não **valem oq o gato enterra na caixa de areia**”. (Significa – valer muito pouco ou nada).

G) numérica: possui, na sua composição léxica, um numeral. Exemplos:

- (15) “Não tem nenhuma explicação de roteiro pra Lurdes não ter **gritado aos 4 ventos** que achou o filho”. (Significado – para todos escutarem, tomarem conhecimento).
- (16) “Carla tinha q ver ela sendo mal falada **nos 4 cantos** da casa pra acordar de vez”. (Significa – em todo lugar).

H) situacional: é empregada em uma situação social precisa ou desencadeada por uma situação específica. Exemplos:

- (17) “Se Thelma **encostar um dedo no fio de cabelo** da Lurdes eu vou na Globo”. (Significa – tocar em alguém mesmo que de forma mínima, é utilizada em situações em que uma pessoa tenta proteger outra).
- (18) “Comunistóides, a **onça vai beber água**! Povo na rua”. (Significa – que algo irá mudar, utilizada em momentos decisivos ou em situações difíceis).
- (19) “Aguardando ansiosa as **cenas dos próximos capítulos** da política nacional”. (Significa – esperar por acontecimentos futuros, utilizadas diante de situações que inspiram curiosidade).

3.5 ANÁLISE DOS DADOS POR EIXO TEMÁTICO

Como já referido anteriormente no primeiro capítulo deste trabalho, as Expressões Idiomáticas possuem como uma de suas características a polilexicalidade que, por sua vez, se refere tanto ao aspecto quantitativo (necessidade de duas ou mais lexias em sua formação estrutural), quanto ao aspecto qualitativo que trata das relações semânticas que se estabelecem entre os itens lexicais que a compõe, dessa forma, significa afirmar que desde o ponto de vista do campo semântico sua significação fixa-se na memória coletiva dos usuários como uma única unidade.

Monteiro-Plantin (2014, p. 87) ressalta que nem toda palavra que possui duas ou mais unidades léxicas é necessariamente uma unidade fraseológica e relembra que para tal é preciso que haja “desvio do sentido literal em pelo menos um dos constituintes*”.* Nessa perspectiva, Zavaglia (2006, p. 29) reitera que estruturalmente as “[...] partes combinatórias não podem ser desmembradas em unidades singulares de sentido. Ao contrário, o significado deve ser depreendido a partir da totalidade da unidade frasal que terá um sentido próprio e peculiar”.

Nessa lógica, pode-se considerar que este aspecto está diretamente relacionado com o elemento cultural das EI, visto que, elas costumam empregar significação partindo da reflexão dos valores culturais da comunidade linguística a que pertencem, reiterando a relação entre léxico e cultura, como pontua Ortiz Alvarez (2013):

> São fórmulas coletivas e tradicionais que refletem a mentalidade de um povo, sua história, seus costumes, crenças e estados afetivos, aos olhos de quem saiba reconhecê-las e investigar a visão de mundo que refletem. Assim, no correr dos séculos, essas fórmulas foram plasmadas em um vasto número de expressões – muitas vezes caracterizadas populares -, que seriam portadoras das vivências de uma ou mais gerações aplicadas no cotidiano. (ORTIZ ALVAREZ, 2013, p. s/n).

Considerando as observações apresentadas propõe-se uma análise dos dados, sob uma ótica semântica, para observar se as expressões utilizadas dentro de um mesmo eixo temático[78] do *corpus* (reality shows, política, novelas) possuem algum ponto de relação entre si, por meio do sentido global aplicado a significação das EI, a partir dos contextos de uso.

3.5.1 Eixo temático de *reality shows*

A) *Hashtags* referentes ao eixo temático:

- #DAYANEMELLOCAMPEA
- #ForaArthur
- #paredaofalso
- #CarlaNoQuartoFalso
- #BBB21Carla
- #CATBBB
- #ForaCarla
- #Camila
- #ForaJuliette

B) Expressões com correlação semântica:

As expressões *agir pelas costas* e *apunhalar ele pelas costas*, referidas no exemplo (1) e (2) fazem parte de um nicho de significação que indica atitudes que foram realizadas para alguém, sem o seu conhecimento. Tais EI costumam ser utilizadas em situações, nas quais, os usuários consideram a possibilidade haver mentiras ou omissões na relação estabelecida. No primeiro caso é um comentário destinado aos participantes do reality BBB21 que falaram mal da atriz Carla Diaz ao acreditarem que essa havia sido eliminada do programa, no entanto, ela participou de dinâmica do programa de paredão falso.

O segundo caso refere-se a estratégia utilizada pela atriz ao retornar da dinâmica. O público acreditava que ao ver as atitudes e falas do companheiro amoroso no jogo, ela teria uma

[78] Nesta dissertação, entende-se que eixo temático é um núcleo de tema do qual podem surgir subtemas relacionados ao cerne da base.

atitude incisiva, contudo ocorreu o contrário, dessa forma houve especulações de que ela estaria se preparando para uma possível vingança.

- (1) "É fácil se dizer arrependido depois de tudo que falou e como **agiu pelas costas** após o retorno da Carla".
- (2) "Pensando aqui... E SE TALVEZ for estratégia dela ficar com ele de novo só pra **apunhalar ele pelas costas?".**

Os exemplos (3), (4) e (5) apresentam as expressões *cair no papinho dele*, *cair nos braços* e *estar caidinha por ele* como sinônimas, referindo-se ao ato de estar tão envolvido amorosamente com alguém, ao ponto de manter uma relação de dependência. Os três contextos dizem respeito ao relacionamento amoroso entre a atriz Carla e o professor Arthur.

- (3) "AMIGA EU NÃO VOU MAIS **CAIR NO PAPINHO DELE**! ".
- (4) "Espero que quando a Carla entrar ela dê um fora no Arthur e **cai nos braços** de Juliette, Camilla e Pocah".
- (5) "Ja a Carla va ficar ouvindo só do Arthur, pq ela **ta caidinha por ele** e magoada, então vai querer saber se ele pensa nela".

As expressões *um ombro amigo* e *chorando no ombro dela* presentes nos exemplos (6) e (7), fazem parte do eixo semântico relacionado a situações onde um indivíduo pede apoio diante de situações difíceis. No caso dos contextos apresentados, os usuários mencionam a diferença de comportamento que a participante do BBB21, Juliette, dispensa aos seus amigos e a que eles dispensam para ela.

- (6) "O povo fala mal da Juliette, mas é Juliette que tá ali, aconselhando, dando apoio, **um ombro amigo**".
- (7) "Arthur falou que está de saco cheio da Juliette, mas na festa estava **chorando no ombro dela".**

Outras duas expressões utilizadas como sinônimos são: *com fogo nos olhos* e *com sangue nos olhos*, utilizadas para fazer referência a pessoas que agem de forma destemida, sem preocupar-se com os obstáculos. Nos exemplos (8) e (9) os usuários pedem a que ao retornar da dinâmica de paredão falso do programa, Carla tenha um maior afinco em suas decisões, utilizando as informações extras que possui.

- (8) "Carla volta **com fogo nos olhos** mulher, pelo amor de Deus. Mete o loko □□□".
- (9) "@Carladiaz Ainda bem que ela será a mais votada e voltará **com sangue nos olhos**! #foraarthur".

As expressões *fogo no parquinho, quebrando o barraco, fazer barraco* e *bater de frente*, presentes nos exemplos (10), (11), (12) e (13) fazem parte do eixo temático situacional que se refere a brigas. Nos contextos em questão, os *tweets* pedem por atitudes mais radicais das participantes do reality, Carla e Juliette.

- (10) “Genteee ela tá esperando a faísca pra começar o **fogo no parquinho**, aguardem a Carla encarnar a Karine”.
- (11) “Gente, vamo acordar, de onde vocês tiraram que a Carla vai voltar **quebrando o barraco**? ”
- (12) “Vai @FreireJuliette esse é o seu momento de **fazer barraco** pelo amor de Deus”.
- (13) “Ele vai ver todo mundo falando mal da Carla e quando voltar, vai **bater de frente** com todos”.

Os exemplos (14) e (15) mostram as expressões *sem sal* e *água de chuchu* referem-se a um campo semântico sentimental de indiferença. Nas duas situações, os comentários referem-se a atuação e possibilidade de saída dos participantes no programa, demonstrando sua insatisfação.

- (14) “Meu Deus que quadrilha esse Projota essa Carla Dias,esse falso nojento Arthur,e essa dorminhoca **sem sal** poka”.
- (15) “Paredão **água de chuchu**!!!!!!”

3.5.2 Eixo temático de política

A) *Hashtags* referentes ao eixo temático:

- #Bolsonaro2022
- #BolsonaroGenocidaSim
- #ConteComigoBolsonaro
- #DoriaGenocida
- #EuSouOExercitoDoBrasil
- #PLdoGasOriginal
- #GENOCIDA
- #Lulaland

B) Expressões com correlação semântica:

As expressões: *passador de pano* e *passar a mão na cabeça* presente nos exemplos (1) e (2) referem-se simultaneamente a situações em que se abstém alguém da responsabilidade sobre algo de forma proposital.

- "Onde está seu tuite sobre a mansão do flavinho seu **passador de pano**".
- "Não **passando a mão na cabeça** do luloso, mas dadas as circuntâncias, na atual conjuntura dos fatos, os que ainda estão do lado genocida e negacionista da história".

Já nos exemplos (3) e (4), as expressões: *barca furada* e *estar no fundo do poço* estão referindo-se à situação do governo atual e no período de 2002, respectivamente e refletem a insatisfação dos usuários com as políticas governamentais aplicadas.

- (3) "Esse governo é **barca furada**".
- (4) "No 2º turno de 2002 **estávamos no fundo do poço**. Agora no melhor dos mundos em 40 anos".

3.5.3 Eixo temático de novelas

A) *Hashtags* referentes ao eixo temático:

- #AForcaDoQuerer
- #AmorDeMae
- #Capitu
- #Thelma
- #NinguemMexeComErnesto
- #Danilo

B) Expressões com correlação semântica:

Nos contextos (1), (2), (3) e (4) são apresentadas quatro expressões: *estar aos pés*, *chegar aos pés, tem que comer muito arroz e feijão* e *tem comer um quilo de fermento* e são utilizadas com o objetivo de expressar a inferioridade de uma pessoa sobre outra. No caso em específico, a comparação é feita entre as personagens Bibi, ex-mulher de Sabiá, e Carine, a atual parceira do traficante, todos referentes a novela "A Força do Querer", da emissora Rede Globo.

- "A cara do Sabiá resume o quanto Carine nunca **estará aos pés** de Bibi perigosa #AForcaDoQuerer".

- “KKKKKKKKKK carine nunca vai **chegar aos pés** da bibi e olha que eu nunca gostei daqueles shows dela”.
- “Carine não chega nem aos pés da Bibi perigosa. **Tem que comer muito arroz e feijão** minha filha #AForçaDoQuerer”.
- “A Carine querendo ser a Bibi! Minha filha, tu **tem que comer um quilo de fermento** pra ter um pouquinho da existência da Bibi”.

As EI *meter o pau, quebrar a cara* e *voar em cima* fazem parte do eixo temático situacional relacionado com brigas. No caso dos exemplos, referem-se as personagens Thelma e Lurdes da novela “Amor de Mãe”, da emissora Rede Globo. Na cena televisionada Lurdes é sequestrada por Thelma, após descobrir que essa última sequestrou o filho da primeira ao nascer.

- (5) “Lurdes do céu, **mete o pau** na Thelma e foge por favor”.
- (6) “eu tô a ponto de invadir essa televisão e **quebrar a cara** da thelma, que raivaaaa”.
- (7) “Lurdes já era pra ter **voado em cima** de Thelma”.

3.6 ANÁLISE DE POSSÍVEIS CANDIDATOS A “NOVAS” EXPRESSÕES IDIOMÁTICAS

Haja vista que a internet é um ambiente onde as informações e novidades se movimentam com muita rapidez e de forma bastante espontânea, a língua encontra campo para frequentes transformações. Sejam elas momentâneas ou duradouras, o fato é que estudar essas mutações fornece material expressivo para os mais variados estudos linguístico-sociais. Tal característica da internet reflete proporcionalmente na rede social escolhida, visto que é um universo informal e frequentado, majoritariamente, por indivíduos jovens, situação favorável para as comentadas mudanças linguísticas.

Com essas observações em mente e pensando nas possibilidades que os dados apresentam para pesquisas futuras, propõe-se aqui, uma demonstração preliminar de parte dos 141.000 blocos de comentários armazenados, nos quais, percebe-se uma maior aparição de candidatos a Expressões Idiomáticas que podem ser considerados relativamente “novos”[79]. São eles: *passar pano, jantar cedo, dar biscoito, estar na Disney, dar PT, fingir demência, estar/ficar pistola, (dar/ter) close errado, estar/ser pago, (colocar) fogo no parquinho.* Vale ressaltar que aproximações que comprovem tal assertiva podem ser obtidas mediante estudo

[79] Realizou-se a consulta das expressões nos dicionários de modalidade online “Aulete Digital” e “Priberam”, que por estarem dispostos na internet tem constante atualização dos dados. O dicionário Aulete não identificou nenhuma das expressões mencionadas, já o dicionário Priberam apresenta registro apenas do termo “ficar pistola”, não reconhecendo a variante com o verbo “estar”.

aprofundado. Um possível campo de estudo é, por exemplo, o neológico que trabalha com a teoria da Lexicologia Social de Matoré (1953). Desse modo, a intenção aqui é uma breve apresentação de um campo vasto para investigações posteriores, e não uma análise efetiva. Na sequência, são trazidos exemplos de comentários onde as expressões citadas aparecem, seguidos de explicação do uso semântico averiguado na rede social.

- (1) "Raíssa brigando pela pizza: Eu automaticamente indo **passar pano** por que eu faria o mesmo, até porque por comida eu sou capaz de fazer uma guerra kakakakakakakakak"
- (2) "eu to AMANDO the penthouse, um drama resumido a personagens lindos, de caráter duvidoso e moralmente controversos pra gente **passar pano**"
- (3) "além do fiuk ser um péssimo ator ainda deram um papel horrível pra ele. se a irene matar o ruy eu vou **passar pano** pra ela sim"

Os três primeiros exemplos mostram a expressão *passar pano*. O primeiro fala sobre Raíssa, uma participante do *reality show* "A Fazenda" exibido no ano de 2020, pela emissora Rede Record. O segundo se refere à série televisiva coreana "The Penthouse: War in Life". Já o terceiro, faz afirmações sobre o ator brasileiro Fiuk e seu personagem – Ruy – contracenando com a personagem da atriz, também brasileira, Débora Falabella – Irene - na novela "A Força do Querer", da emissora Rede Globo. De forma denotativa, passar pano faz referência à ação de limpar algum tipo de superfície com o auxílio de material têxtil. No *corpus* constituído, os contextos nos quais a expressão aparece refletem o sentido conotativo de ser favorável/defender uma pessoa que comete ou cometeu erros.

- (4) "La vai a Juliette **jantar cedo** mais uma vez"
- (5) "HAHAHAHA eu sabia que ela ia **jantar cedo** hoje! RT @Anitta Ao invés de trabalhar ficam de gracinha no *Twitter*... esse é nosso governo teen... até eu to mais ocupada fazendo algo pelo meu país do que você, meu querido"
- (6) "Cheguei agora mas depois de ouvir essa pergunta do "conheço um hospital que funcionou" eu só espero que o mandetta esteja com fome para **jantar cedo**. Pq qq medico jantaria esse senador."

Nos comentários quatro, cinco e seis, observa-se a expressão *jantar cedo*. No primeiro, o usuário dá sua opinião sobre uma atitude de Juliette, participante do *reality show* "Big Brother Brasil" exibido no ano de 2021, pela emissora Rede Globo. O segundo é um comentário sobre a cantora brasileira Anitta, em meio a uma discussão com o atual ministro do Meio Ambiente

Ricardo Salles. Já o terceiro é um comentário sobre a expectativa de discurso do ex-ministro da saúde Luiz Henrique Mandetta na CPI Covid[80], no ano de 2021. Como já mencionado no primeiro tópico do capítulo 2, o conjunto lexical em questão não está relacionado à uma refeição realizada prematuramente. Os contextos explicitados nos comentários demonstram que *jantar cedo* tem sentido figurado e é usado quando um indivíduo expõe sua opinião utilizando argumentos tão pertinentes que deixa o receptor sem resposta ou reação.

- (7) "Eu vou sim **dar biscoito** pro meu cachorro, pq ele eh lindo!!"
- (8) "Nem sei se alguém vai me **dar biscoito**, mas hoje tô biscoiteira sim"
- (9) "Eu só acho que os cactos precisam parar de **dar biscoito** a verificado q só tão aproveitando esse momento de engajamento e ir votar e também parar de responder torcida q n vota e faz o fav perder a final. não queiram ser os vigorentos e no último momento fazer a ju perder o prêmio, votem."

O sétimo comentário, assim como o oitavo e nono, traz a expressão *dar biscoito*. Comumente essa unidade é encontrada na internet com o sentido conotativo relacionado a beleza e/ou atenção. Isto é, *dar biscoito* não tem relação com a ação de fornecer tal produto alimentício, mas sim, espalhar elogios/atenção a determinada pessoa, ou, como pode ser observado no comentário sete, animal que seja considerado pelo enunciador como possuidor de encanto, graça, formosura, etc, ou esteja carente.

- (10) "Deve **estar na Disney** o Aguero"
- (11) "todo calouro acha que **ta na Disney** quando entra no cefet.eu me incluo"
- (12) "Quer me beijar e não responde meus stories, ta achando que **ta na Disney** 9vinho?"

Os comentários dez, onze e doze apresentam a expressão *estar na Disney*. O primeiro da seleção se refere ao futebolista argentino Leonel Agüero Del Castillo, o segundo aos calouros do CEFET[81], e o terceiro a uma pessoa que aparentemente gostaria que o rapaz por quem está interessada – "9vinho", abreviação de "novinho"[82] – lhe desse atenção nos *stories*, ferramenta da rede social *Instagram*. Em contexto não figurado estar na Disney significaria o óbvio, estar

[80] CPI ou Comissão Parlamentar de Inquérito é uma investigação, conduzida pelo Poder Legislativo, que transforma a casa parlamentar em comissão para ouvir depoimentos e tomar informações diretamente.
[81] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca.
[82] Novinho/a é um termo comumente utilizado pelo público jovem para se referir a alguém por quem nutre interesse amoroso/ sexual.

presente nos parques do complexo *Walt Disney World*[83]. Porém, nos comentários encontrados, a expressão aparece com sentido figurado utilizado quando alguém pensa que a realidade é utópica e mágica como a fantasia criada por Walt Disney.

- (13) “PRECISO DANÇAR, PRECISO BEIJAR, PRECISO **DAR PT** E ME ENTORPECER, PRECISO DE FEEEEEEEEEEEEEEEESTA. SOCORRO ESTOU COLAPSANDO”
- (14) “Prometi que n ia **dar pt** e não dei, coisa boa acordar sem ressaca”
- (15) “Não posso **dar pt** que a Mimi fica jogando na minha cara que deu banho em mim, credo em”

Já os comentários treze, quatorze e quinze, mostram a expressão *dar PT*. A sigla “PT” faz referência à “perda total”, unidade usada quando um veículo automobilístico tem sua estrutura tão prejudicada devido a acidente, que não possui conserto viável. Entretanto, no material analisado a expressão *dar PT* tem sentido conotativo para situações em que um indivíduo faz uso de bebida alcoólica em grande quantidade, causando danos físicos posteriores como, por exemplo, amnésia momentânea e ressaca no dia posterior.

- (16) “eu odeio gente que tenta se sair por cima de tudo, gente que se faz de vítima p tentar sair como coitadinhx. e eu tô tentando o MÁXIMO **fingir demência** e não perder a cabeça c isso, porque eu odeio ver gente falando o que não sabe.”
- (17) “a gente pode até **fingir demência** mas uma hora ou outra tem que encarar a realidade né”
- (18) “Nao suporto ver nada sobre o Paulo hoje, nao quero acreditar, vou **fingir demência**, esse dia não aconteceu”

Os comentários dezesseis, dezessete e dezoito trazem a expressão *fingir demência*. Como é possível inferir, a expressão é figurada e reflete uma situação na qual a pessoa finge não estar vendo, compreendendo ou participando de determinada circunstância por considerá-la revoltante, vergonhosa, inviável, inadmissível ou triste. Exemplo desse último caso é o comentário dezoito que faz referência a morte do ator brasileiro Paulo Gustavo.

- (19) “Ana Clara deve **tá pistola** com o Tiago que contou os seguidores primeiro que ela”

[83] Complexo de parques fundado por Walt Disney e Roy O. Disney, e localizado na Flórida, nos Estados Unidos da América.

- (20) “Hj vou participar da aula,agr é vdd pq minha mãe **tá pistola** cmg”
- (21) “Minha mãe **tá pistola** comigo pq eu to na cozinha fazendo café e barulho kkkkk”

Estar pistola, por sua vez, não tem relação com artefatos bélicos. Os comentários dezenove, vinte e vinte um mostram ocasiões nas quais a pessoa envolvida se mostra brava, nervosa, etc. Enquanto os dois últimos exemplos trazem situações com mães e filhos, o primeiro faz referência a dois apresentadores de programa de televisão, Ana Clara Lima e Tiago Leifert. O contexto geral é, novamente o *reality show* “Big Brother Brasil” do ano de 2021.

- (22) “o problema de obra palestrinha demais é que 99% das vezes **tem close errado** no meio”
- (23) “vim no mundo pra **da close errado**”
- (24) “Tô bem feliz por ele se sentir livre pra ser quem é, espero que continue assim e seja muito feliz, mas a bicha só **deu close errado** kkkkk”

Os comentários vinte e dois, vinte e três e vinte e quatro, mostram a expressão *ter/dar close errado*. Como empréstimo da língua inglesa, o termo *close* pode ser traduzido de muitas maneiras, mas o que origina a expressão é a redução de *close-up,* enquadramento fechado de fotografia ou vídeo, mostrando apenas uma parte do todo, geralmente o rosto de uma pessoa. *Ter/dar close errado* não é literalmente errar no enquadramento. De maneira conotativa a expressão indica que um indivíduo agiu de maneira equivocada em determinada situação.

- (25) “Fazia um dia inteiro que eu não chorava de amor pelo Xiao Zhan, mas agora posso dizer que: O de hoje **tá pago**.”
- (26) “Treino de hoje, **foi pago** p hoje e amanhã”
- (27) “O treino de segunda **ta pago**. bom diaaaa”

Estar/ser pago, de forma denotativa, significa realizar o pagamento de estabelecido valor financeiro. Entretanto, os exemplos mostram a expressão com sentido conotativo ao significar que determinado compromisso foi cumprido. A expressão é comumente usada em contextos que englobam exercícios/treinamentos físicos, mas o primeiro exemplo mostra que ela também pode estar presente em outros cenários.

- (28) “GIL DO VIGOR, você entregou entretenimento pra gente e muito **fogo no parquinho**. O Brasil tá lascado, mas te ama...”
- (29) “BBB 2022 podia ser com bookstan cara. Quer ver **fogo no parquinho**, boninho? Pega as bookstan”

- (30) “eu e a Carol **colocando fogo no parquinho** de maneira delicaderrima KKKKKKKKKA amo”

Os últimos exemplos apresentam a expressão *(colocar) fogo no parquinho*. Os comentários vinte e oito e vinte e nove fazem novamente referência ao *reality show* “Big Brother Brasil” do ano de 2021, já o trinta traz contexto pessoal. Todos usam a expressão com sentido figurado, o que não significa que fogo está literalmente sendo colocado em um parque, mas sim que, utilizando outra expressão idiomática, alguém está *colocando lenha na fogueira*.

Diante das observações teóricas e metodológicas apresentadas neste trabalho, tal qual a análise de dados quantitativa e qualitativa realizadas, no próximo tópico serão explicitadas as considerações finais acerca da pesquisa, pautando-se tanto nos aspectos positivos encontrados, quanto nos desafios que foram travados ao longo da investigação.

CONSIDERAÇÕES FINAIS

Esta dissertação objetivou a elaboração de um *corpus* linguístico, constituído por meio de comentários, sobre os assuntos mais debatidos, em período específico, na rede social *Twitter*, verificando ferramentas informáticas que auxiliassem nessa composição. De modo específico, pretendeu-se: a) elaborar códigos e matrizes computacionais que funcionassem como ferramentas virtuais para a coleta diária dos comentários dispostos nos *Trending Topics*; b) promover a discussão sobre a produtividade de um *corpus* de comentários de redes sociais, no que diz respeito à uma amostra de ocorrência de expressões idiomáticas, considerando a hipótese de que os usuários, no geral, se comunicam por meio de uma linguagem mais coloquial; c) Contribuir com o processo de consolidação dos estudos no âmbito da Lexicologia e da Fraseologia, especialmente na relação que traçam com a área da Informática.

Compreende-se, aqui, que um embasamento teórico conciso, claro, objetivo e correlacionado é essencial para a constituição de uma base de dados confiável. Portanto, partindo dos objetivos elencados, o primeiro capítulo deste trabalho foi estruturado a fim de discutir os pressupostos teóricos da Lexicologia, da Fraseologia e da Linguística Computacional, tanto no que se refere a Linguística de *Corpus*, quanto, mais especificamente ao tratamento automático da língua. Buscando, assim, além de subsidiar o objetivo geral, demonstrar a correlação existente entre as áreas de estudo, sobretudo, no que diz respeito à aplicação prática da Informática e da Lexicologia.

No segundo capítulo, por sua vez, almejou-se a apresentação de um percurso histórico sobre a criação da rede social *Twitter*, fonte do *corpus* desta pesquisa, e as particularidades que circundam esse universo midiático-social. Ademais, foram expostas as ferramentas utilizadas na constituição do banco de dados, explanação de fenômenos técnicos relacionados ao estudo - como preceitos sobre linguagens de programação, ambientes virtuais nos quais essa manipulação pode ser realizada, tópicos importantes sobre segurança cibernética, *softwares* auxiliares na coleta dos dados -, e, por, fim, a apresentação dos procedimentos metodológicos para constituição do *corpus*.

Pautando-se na representatividade e extensão apresentada pelos assuntos comentados para a composição do *corpus* optou-se, primeiramente, pela coleta dos *tweets* que faziam parte do assunto de número um nos *Trending Topics*, ou seja, mais comentados pelos usuários. Visando, dessa forma, um banco de dados que ilustrasse o uso da língua de forma natural e

diversificada, adotou-se o método de coletas diárias, recolhendo um total de mil comentários diários, reforçando novamente os critérios de extensão estipulado pela Linguística de *Corpus*.

Ao longo do início da coleta, porém, foi possível perceber o aparecimento frequente de "#" estrangeiras no topo dos *Trending Topics*, e nomes próprios com significado global – como "Jesus" e "Corinthians", por exemplo. Tal situação impossibilitou a utilização dessas *hashtags* por estarem ligadas a comentários em línguas diversas que não o português brasileiro. Dessa forma, prioritariamente, decidiu-se pela coleta dos *tweets* que faziam parte do assunto de número um nos *Trending Topics.* Caso esses apresentassem relações estrangeiras em termos de língua, selecionou-se o segundo, terceiro, e assim por diante.

Ponderando os aspectos metodológicos da investigação, é possível observar que as ferramentas apresentadas otimizaram sobremaneira a constituição do *corpus*, haja vista que se a coleta fosse realizada de forma manual, recolher 1000 comentários diários levaria um tempo expressivamente maior. É possível afirmar, então, que usar a linguagem de programação apresenta novas perspectivas e possibilita a criação de *corpora* de maneira ágil e sem dependência de programas pagos e de difícil acesso. Outro aspecto positivo a ser pontuado, foi o uso de *websites* que mostravam as # presentes nos *Trending Topics* nas últimas 24 horas, em diversos países, apresentando inclusive, a possibilidade de limitar a busca por estados e/ou cidades.

Por sua vez, a viabilidade de criação facilitada de *corpora* amplia as possibilidades dos estudos lexicais como um todo. Isto, porque permite que diferentes pesquisas baseadas em *corpus* sejam planejadas e realizadas de maneira dinâmica, sem precisar depender de *corpora* pré-existentes que não consideram, em grande parte, os novos meios de interação social.

Por fim, no terceiro capítulo foram propostas discussões sobre: (a) o uso de *softwares* aliados ao tratamento de lexias complexas e uma breve justificativa da recusa em usá-las nesta pesquisa; (b) "padrões comportamentais" nos assuntos mais comentados listados nos *Trending Topics* do *Twitter;* (c) análise dos dados sob quatros vieses: i) perfil quantitativo; ii) aspectos tipológicos por natureza estrutural e casos especiais, propostos por Xatara (1998); iii) por eixo temático (reality shows, política e novelas); e, iv) uma análise de possíveis candidatos a "novas" EI, elencando a possibilidade de continuação dos estudos.

Em relação aos dados coletados na pesquisa, em termos de origem e propósito, o *corpus* mostrou-se autêntico e produtivo, uma vez que, o site de mídia-social *Twitter* apresenta discursos de usos reais da língua, muito próximos da fala e, que por sua vez, contém em sua composição gírias, expressões idiomáticas, provérbios, dentre outros. Comprovando dessa

forma, a hipótese de que a linguagem coloquial ali apresentada pré-dispõe o ambiente para uso de EI.

Além dos pontos positivos mencionados, também, encontrou-se alguns desafios no percurso, sobretudo no que se refere a leitura e análise dos textos. Por considerar a vastidão do *corpus* textual levantado para a base de dados de suma importância, sua extensão somada ao tempo hábil para realização da pesquisa tornou-se uma dificuldade, haja vista que cada dia de coleta apresentava uma extensão total entre mil e quinhentas e três mil linhas, a depender da quantidade de caracteres utilizada pelo usuário.

Dessa forma, propôs-se um recorte do *corpus* para análise e averiguação da hipótese exposta nos objetivos específicos desta dissertação dentro do tempo estipulado pelo programa de pós-graduação. A amostra de produtividade dos dados foi constituída, então, por trinta e um dia de coleta de comentários, referentes ao mês de março de 2021.

A análise quantitativa dos dados revelou um elevado índice de recorrência das EI em comentários que tinham como tema *reality shows* e política, compreendendo 46% e 37% do *corpus*, respectivamente. São # relacionadas a *reality shows*:

- #BBB21Carla
- #Camila
- #CarlaNoQuartoFalso
- #CATBBB
- #DAYANEMELLOCAMPEA
- #ForaArthur
- #ForaCarla
- #ForaJuliette
- #paredaofalso

São # relacionadas a política:

- #Bolsonaro2022
- #BolsonaroGenocidaSim
- #ConteComigoBolsonaro
- #DoriaGenocida
- #EuSouOExercitoDoBrasil
- #GENOCIDA
- #Lulaland

- #PLdoGasOriginal

Ressaltou-se, ainda, que 94,8% das expressões possuem caráter somático na sua formação estrutural. São exemplos de Expressões Idiomáticas de caráter somático:

- *Abrir o olho*
- *Com fogo nos olhos*
- *Lambendo a bunda*
- *Lágrimas de crocodilo*
- *Passar pela cabeça*
- *Sangue nos olhos*

Ademais, foi possível comprovar que as EI que pertenciam aos mesmos eixos temáticos, possuíam inter-relações semânticas, sendo algumas utilizadas em um mesmo contexto situacional, ou como sinônimas. São exemplos:

- *Agir pelas costas*; *Apunhalar ele pelas costas*
- *Cair no papinho dele*; *Cair nos braços*; *Estar caidinha por ele*
- *Um ombro amigo*; *Chorando no ombro dela*
- *Com fogo nos olhos*; *Com sangue nos olhos*
- *Barca furada*; *Estar no fundo do poço*
- *Estar aos pés*; *Chegar aos pés; Tem que comer muito arroz e feijão; Tem comer um quilo de fermento*
- *Meter o pau*; *Quebrar a cara*; *Voar em cima*

Além disso, é possível comprovar que o *corpus*, em questão, formado por comentários de rede social é produtivo, também, no que diz respeito à aparição de candidatos a "novas" expressões idiomáticas, isso é, ainda não dicionarizadas, como demonstrado no subtópico "3.6 Análise de possíveis candidatos a "novas" expressões idiomáticas". São eles:

- *(Colocar) fogo no parquinho*
- *(Dar/ter) close errado*
- *Dar biscoito*
- *Dar pt*
- *Estar na disney*
- *Estar/ficar pistola*

- *Estar/ser pago*
- *Fingir demência*
- *Jantar cedo*
- *Passar pano*

Vale ressaltar que as discussões apresentadas sobre esses candidatos não são análises aprofundadas, mas sim apresentações que explicitam um campo vasto para investigações posteriores, principalmente as neológicas que trabalhem com a teoria da Lexicologia Social de Matoré (1953).

Por fim, diante do exposto nos capítulos desta dissertação, acredita-se que o resultado esperado foi alcançado, visto que, todos os objetivos foram cumpridos, apresentando como produto desta pesquisa o desenvolvimento dos códigos necessários para a coleta de comentários do *Twitter*, no período de dezembro de 2020 a maio de 2021. Da mesma maneira que, a comprovação da hipótese de alto índice de ocorrência de Expressões Idiomáticas no discurso dos usuários dessa rede, constando uma amostra com o total de 232 EI, no intervalo temporal de 31 dias, referente ao mês de março de 2021.

Dita situação corrobora com os preceitos que afirmam que a internet é um ambiente onde as informações e novidades se movimentam com muita rapidez e de forma bastante espontânea, e que esse grupo de unidades fraseológicas está diretamente ligado ao caráter informal e descontraído que os textos de rede social carregam.

Com base nesses apontamentos, pode-se afirmar que a investigação, de maneira geral, alcança evidência por ser inovadora e produtiva. Em vista disso, conclui-se que esta discussão, e o desenvolvimento do estudo como um todo, contribuem sobremaneira para o avanço das pesquisas pautadas em Fraseologia e Linguística Computacional, escrevendo um novo capítulo sobre as relações entre Linguística e Informática.

REFERÊNCIAS BIBLIOGRÁFICAS

A chegada. Direção: Denis Villeneuve. Produção de Shawn Levy, Dan Levine, Aaron Ryder e David Linde. Estados Unidos: Paramount Pictures, 2016. 1 DVD (116 min.).

ANDRADE, Mário de. **Poesias Completas**. Belo Horizonte: Villa Rica, 1993.

AULETE DIGITAL. **Dicionário da Língua Portuguesa na Internet**. Rio de Janeiro: Lexikon, 2021. Disponível em: <https://aulete.com.br/>. Acesso em: 26 de março de 2021.

BAKHTIN, M.; VOLOSHINOV, V. N. **Marxismo e Filosofia da Linguagem**. São Paulo: Editora Hucitec, 1988.

BARBOSA, M. A. Lexicologia, Lexicografia, Terminografia: objeto, métodos, campos de atuação e de cooperação. Estudos Linguísticos. Franca: Unigran/GEL, 1991.

BARREIROS, L. L. S. O uso de ferramentas computacionais na elaboração do Vocabulário de Eulálio Motta AntConc e FLEx. In: **A Cor das Letras**. Feira de Santana: UEFS, v. 18, n. 2, 2017.

BÍBLIA. **Bíblia Sagrada**. São Paulo: Paulinas, 2005.

BIDERMAN, M. T. C. **Teoria linguística**: linguística quantitativa e computacional. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

_____. A Estrutura Mental do Léxico. In: **Filologia e Lingüística**. São Paulo: Edusp, v. 1, 1981.

_____. Léxico e vocabulário fundamental. In **Alfa**: Revista de lingüística. São Paulo, Universidade Estadual Paulista, v. 40, 1996, p. 27-46.

_____. Dimensões da Palavra. In: **Filologia e Lingüística Portuguesa**. São Paulo: Humanitas, n. 2, 1998.

_____. O Conceito Linguístico de Palavra. In: BASÍLIO, M. (org.). **A Delimitação de Unidades Lexicais**. Rio de Janeiro: Grypho, v. 1, 1999.

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

_____. Linguística de *corpus*: Histórico e problemática. In: **D.E.L.T.A**., v. 16, nº 2, 2000.

BIG DATA: O que é, conceito e definição. Cetax, 2020. Disponível em: <https://www.cetax.com.br/blog/big-data/>. Acesso em: 26 de março de 2021.

BLOG PLANETA MATEMÁTICA. **Homepage**. Disponível em: < http://planeta-matematica.blogspot.com/>. Acesso em: 31 de março de 2021.

BORBA, Francisco da Silva. **Dicionário Unesp do português contemporâneo**. São Paulo: UNESP, 2004.

BOYER, C. B.; MERZBACH, U. C. **História da Matemática**. Tradução: Helena de Castro. São Paulo: Blucher, 2012.

BRASIL. **Constituição da República Federativa do Brasil**. Brasília, DF: Senado Federal: Centro Gráfico, 1988.

BYRNE, R. **O Segredo**. Rio de Janeiro: Sextante, 2015.

CORI, M. Des méthodes de traitement automatique aux linguistiques fondées sur les *corpus*. In: **Langages, 171**. 10.3917/lang.171.0095, 2008.

CORI, M.; MARANDIN, J.-M. La linguistique au contact de l'informatique: de la construction des grammaires aux grammaires de construction. In: **Histoire Épistémologie Langage**. SHESL/EDP Sciences, 23, pp.49-79, 2001.

CORPAS PASTOR, G. **Manual de fraseología española**. Madrid: Gredos, 1996.

COSERIU, E. Introducción al estudio estructural del léxico. In: **Principios de semántica estructural**. Madrid: Gredos, 1977.

DADO. In*:* **AULETE DIGITAL, Dicionário da Língua Portuguesa na Internet**. Rio de Janeiro: Lexikon, 2021. Disponível em: <https://aulete.com.br/dado>. Acesso em: 26 de março de 2021.

DAWKINS, R. **The Selfish Gene**. New York: Oxford University Press, 1976, 2007.

EUZÉBIO, N. G. S. F. **Protótipo de dicionário monolíngue de expressões idiomáticas somáticas do português.** Dissertação (Mestrado em Estudos de Linguagem) – Faculdade De Artes Comunicação E Letras, Universidade Federal de Mato Grosso do Sul, 2021.

FERREIRA, A. B. H. **Dicionário da língua portuguesa**. 5. ed. Curitiba: Positivo, 2010.

HOUAISS, A. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro, Ed. Objetiva, 2001.

FINATTO, M. J. B; REBECHI, R. R; SARMENTO, S; BOCORNY, A. E. P. (Orgs.) **Linguística de corpus:** perspectivas [recurso eletrônico]. Porto Alegre: Instituto de Letras - UFRGS, 2018.

INMOMENT. **Homepage**. Disponível em: <https://inmoment.com/>. Acesso em: 26 de março de 2021.

KLARE, J. Lexicologia e fraseologia no português moderno. In: **Revista de Filologia Românica, IV.** Madrid: Editorial de la Universidad Complutense, 1986.

KOTLER, P. **Marketing para o século XXI:** Como criar, conquistar e dominar mercados. São Paulo: Futura, 1999.

LIAISE. **Homepage**. Disponível em: < https://liaise.com.br/>. Acesso em: 26 de março de 2021.

LOUBAK, A. L. O que significa "pprt"? Veja 13 gírias e expressões usadas no *Twitter*. **Techtudo**, 2020. Disponível em: <https://www.techtudo.com.br/listas/2020/05/o-que-significa-pprt-veja-13-girias-e-expressoes-usadas-no-twitter.ghtml>. Acesso em: 23 de março de 2020.

LUQUE DURÁN, J. D; MANJÓN POZAS, F. J. Tipología léxica y tipología fraseológica: universales y particulares. In: PAMIES BERTRAN, A; LUQUE DURÁN, J. D (orgs.). **Léxico y Fraseología**. Granada: Método, 1998.

MARQUES, E. A. Análisis cognitivo-constrastivo de locuciones somáticas del español y del portugués. Tesis Doctoral (Lingüistica Aplicada) – **Departamento de** Filología, Universidad de Alcalá, 2007.

MATORÉ, G. **La méthode en lexicologie: domaine français**. Paris: Didier, 1953.

MELLADO BLANCO, C. **Fraseologismos somáticos del alemán**: un estúdio léxico-semántico. Laussane: Peter Lang, 2004.

MICHAELIS. **Dicionário Online de Língua Portuguesa**. São Paulo: Melhoramentos, 2021. Disponível em: < https://michaelis.uol.com.br/>. Acesso em: 26 de março de 2021.

MILLAN, Marília Pereira Bueno. Reality shows: uma abordagem psicossocial. In: **Psicologia, ciência e profissão**. Brasília, v. 26, n. 2, p. 190-197, jun. 2006.

MONTEIRO-PLATIN, R. S. **Fraseologia:** Era uma vez um Patinho Feio no Ensino da Língua Materna. Ceará: Editora da Universidade Federal do Ceará, 2014.

MONTORO DEL ARCO, E. T. Clasificaciones de las Ufs: el lugar de las locuciones. In: _____. **Teoría fraseológica de las locuciones particulares. Las locuciones prepositivas, conjuntivas y marcadoras del español**. Frankfurt am Main: Peter Lang, 2006.

MUSSO, P. Sociedade Midiatizada. In: MORAES, Dênis de (Org.). **Ciberespaço, figura reticular da utopia tecnológica**. Rio de Janeiro: Mauad X, 2006

O jogo da imitação. Direção: Morten Tyldum. Produção de Nora Grossman, Ido Ostrowsky e Teddy Schwarzman. Estados Unidos; Reino unido: StudioCanal; The Weinstein Company, 2014. 1 DVD (114 min.).

ORENHA, A.; CAMARGO, D. C. A extração de unidades fraseológicas especializadas a partir de *corpora* paralelos e comparáveis. In: **The ESPecialist**, vol. 30, nº 1, 2009.

OTHERO, G. A.; MENUZZI, S. M. **Lingüística Computacional**: teoria & pratica. São Paulo: Parábola Editorial, 2005.

PETTER, M. Linguagem, língua, linguística. In: FIORIN, J. L. (org.). **Introdução à lingüística I:** Objetos teóricos. São Paulo: Contexto, 2002

POTTIER, Bernard. **Lingúística geral:** teoria geral e descrição. Trad. de Walmírio Macedo. Rio de Janeiro: Presença, 1978.

PRIBERAM. **Dicionário Priberam da Língua Portuguesa Online.** Lisboa: Priberam Informática, 2021. Disponível em: < https://dicionario.priberam.org/>. Acesso em: 26 de março de 2021.

ROSA, N. Jovens dominam instagram com prints de memes do *Twitter*. **Canaltech**, 2020. Disponível em: <https://canaltech.com.br/redes-sociais/jovens-dominam-instagram-com-prints-de-memes-do-twitter-116394/>. Acesso em: 23 de março de 2020.

RUIZ GURILLO, L. Aspectos de fraseologia teórica española. In: **Cuadernos de Filología**. Valencia: Artes Gráficas, 1997.

SAUSSURE, F. **Curso de lingüística geral**. São Paulo: Cultrix, 1969.

SILVA, D. B. **Redes sociais e virtuais:** Um estudo da formação, comunicação e ação social. Dissertação (Mestrado em Design e Arquitetura) – Faculdade de Arquitetura e Urbanismo, Universidade de São Paulo, 2011.

SINCLAIR, J. M. ***Corpus*, concordance, collocation**. Oxford: Oxford University Press, 1991.

TAGNIN, S. E. O. Linguística de *corpus* e Fraseologia: uma feita para a outra. In: ALVAREZ, M. L. O.; UNTERNBÄUMEN, E. H. (Orgs.). **Uma (re)visão da teoria e da pesquisa fraseológicas**. Campinas: Pontes, 2011.

_____; FINATTO, M. J. Bocorny; FROMM, G. Linguística de Corpus: conquistas e desafios. In: Revista de Estudos da Linguagem, [S.l.], v. 29, n. 2, p. 661 – 671, 2021.

TRISTÁ PÉREZ, A. Mª. "Teoría fraseológica: visón general del problema. In: TRISTÁ PÉREZ, A. Mª. **Fraseología y contexto**. La Habana: Ciencias Sociales, 1988.

VIANA, A. Geração dos Millenials: Onde vivem, como pensam, como compram e como vendem. ***Reev***, 2020. Disponível em: < https://reev.co/geracao-dos-millennials/>. Acesso em: 23 de março de 2020.

VIEIRA, R. **Lingüística computacional**: fazendo uso do conhecimento da língua. Entrelinhas, ano 2, n. 4, São Leopoldo: UNISINOS, 2002.

VIEIRA, R.; LIMA, V. L. S. Lingüística computacional: princípios e aplicações. In: **IX Escola de Informática da SBC-Sul.** Luciana Nedel (Ed.) Passo Fundo, Maringá, São José. SBC-Sul, 2001.

WAZLAWICK, R. S. **História da Computação**. Rio de Janeiro: Elsevier, 2016.

WELKER, H. A. Colocações e Expressões Idiomáticas em dicionários gerais. In: **Uma (re)visão da teoria e da pesquisa fraseológicas**. ALVAREZ, M. L. O; UNTERNBÄUMEN, E. H. (Orgs.). Campinas: Pontes, 2011.

XATARA, C. M. O campo minado das expressões idiomáticas. In: **Alfa**. São Paulo, 42 (n. esp.), 1998, pp. 147-159.

XATARA, C. M. Tipologia das expressões idiomáticas. In: **Alfa**. n. 42. São Paulo, 1998b. p. 169-176.

XATARA, C. M.; PARREIRA, M. C. A elaboração de um dicionário fraseológico. In: ORTIZ, A. M. L. e UNTERNBAUMEN, Enrique Huelva (Orgs.). **Uma (Re)Visão da teoria e da pesquisa fraseológicas**. Campinas: Pontes Editores, 2011. p. 69-75.

ZAVAGLIA, C. Lexicologia: o que há por trás do estudo das palavras? In: GONÇALVES, A.V; GÓIS, M. L. S. (Orgs.). **Ciência da linguagem:** o fazer cientifico? 1ed. Campinas, São Paulo: Mercado de Letras, 2012.

ZULUAGA OSPINA, A. **Introducción al estudio de las expresiones fijas**. Max Hueber, Verlag, Tübingen, 1980.

APÊNDICES

Quadro 10: Lista de Expressões Idiomáticas coletadas no mês de março/2021.

Expressões Idiomáticas
Água de chuchu
A batata dele está fritando
A cara nem arde
A cereja do bolo
A cobra morde o próprio rabo
A onça vai beber água
A voz do povo é a voz de Deus
Abaixar a cabeça
Abrir a boca
Abrir mão
Abrir o jogo
Abrir o olho
Agir pelas costas
Aguardar as cenas dos próximos capítulos
Andar de cabeça erguida
Apertarem o cinto
Apunhalar pelas costas
Aquecer o coração
Arrebentar a corda
Atingir o topo
Baixa a bola
Bater de frente
Bater em cachorro morto
Borrar as calças
Botar pressão
Brincaram com fogo
Cagada atrás de cagada
Cair na onda
Cair no papinho
Cair nos braços
Cara de cu
Cara de pau
Cavar a própria cova
Chegar aos pés
Chegar no ouvido
Chorando no ombro dela
Chupa essa manga
Chutar cachorro morto
Colocar a mão na massa
Colocar a mão no fogo
Colocar as cartas na mesa
Colocar as coisas no eixo

Colocar comida na mesa
Colocar na balança
Colocar o rabo entre as pernas
Com fogo nos olhos
Com o cu na mão
Com o pé nas costas
Com olhos abertos
Com sangue nos olhos
Com todas as forças
Comendo o cérebro
Comer muito arroz e feijão
Comer o cu
Comer os dedos
Conversa para boi dormir
Correr atrás
Dando corda
Dar a volta por cima
Dar as caras
Dar bola dentro
Dar de cara
Dar gás
Dar o golpe do baú
Dar pitaco
Dar seu sangue
Dar um tapa
Dar uma mãozinha
De boca aberta
De braços cruzados
De meia tigela
Dedo duro
Deixar a máscara cair
Deixar a própria sorte
Deixar nas mãos
Deixar no ar
Desenhar para todos
Dobrar a língua
É cadelinha
É uma bolha
Encostar um dedo em um fio de cabelo
Engolir o choro
Entrar na sua onda
Entre a cruz e a cadeirinha
Entregar as ovelhas ao lobo
Ergue a cabeça
Esfregar na cara
Esperar o cabaré pegar fogo

Está caidinha por ele
Está pegando fogo
Estamos no fogo
Estar aos pés
Estar às minguas
Estar cagando
Estar cagando e andando
Estar com o coração partido
Estar de saco cheio
Estar mais sujo que pau de galinheiro
Estar na boca de alguém
Estar na cara
Estar nas suas mãos
Estar no fundo do poço
Estar roendo as unhas
Enxergar o próprio umbigo
Falar até papagaio fala.
Farinha do mesmo saco
Faz de gato e sapato
Fazer a lição de casa
Fazer cu doce
Fazer os diabos
Fazer um barraco
Ficar alfinetando
Ficar babando alguém
Ficar babando ovos
Ficar de boca fechada
Ficar na mão de
Fogo no Parquinho
Foi na onda
Fora da bolha
Gastar meu réu primário
Girar em torno do umbigo
Gritar aos quatro ventos
Jogar num lamaçal
Jogar sujo
Jogar tudo no chão
Lágrimas de crocodilo
Lambendo a bunda
Lambendo o cu
Lamber o rabo
Lamber pau
Largar a mão
Lavando roupa
Levar chifre
Levar o pão para dentro de casa

Mais claro que a água
Mala sem alça
Mandar no pedaço
Mãos sujas de sangue
Máscaras caindo
Meia boca
Meter a boca
Meter o pau
Meter o pé
Morrer na minha mão
Não largar o osso
Não ter pé nem cabeça
Não usar o cérebro
Não vai colar
Não valer o que o gato enterra
Nas mãos
Nem fede nem cheira
Nem que meu dedo tenha que cair
No chão
No muro
Nos 4 cantos
O choro é livre
O golpe tá aí, cai quem quer
O pai tá on
O pau vai torar
Ombro amigo
Pagar um preço
Passador de pano
Passar a mão
Passar a mão na cabeça
Passar pano
Passar pela cabeça
Pegar a visão
Pegar no pulo
Pegar pra Cristo
Peidar fino
Peidar na farofa
Pelas costas
Perdi o chão
Ponto fraco
Por uma luz na cabeça
Precisa cair
Preparar o lombo
Preparem o coração
Provando do próprio veneno
Quebrando o barraco

Quebrar a cara
Rachar de rir
Rasgar as cortinas
Rasgar o cu
Refrescar a cabeça
Sair por cima
Sangue de barata
Sangue nos olhos
Sem cérebro
Sem conhecer a peça
Sem nem abrir a boca
Sem sal
Sentar o dedo
Sentir o gosto do próprio veneno
Separar os homens dos meninos
Ser a vaquinha do presépio dele
Ser cachorrinho
Ser capacho
Ser duro
Ser feita de açúcar
Ser muito areia
Ser planta
Ser uma barca furada
Será a última pá no enterro dele
Serviu muito
Só de olho
Soltar a mão
Subir a cabeça
Tava na Cara
Te jogar em um poço
Tem que cair
Tem que comer um quilo de fermento
Tentar derrubar
Ter a batata assando
Ter as mãos sujas de sangue
Ter coração explodindo
Ter na mão
Ter sangue nas mãos
Ter um pingo
Tirar as maçãs podres
Tirar o cavalinho da chuva
Tocar o terror
Vai lavar uma louça!
Valer a pena
Vamos biscoitar
Vamos em peso

Ver espumando
Virar o jogo
Virar uma barata
Voar em cima
Voltar a pastar

Fonte: Elaborada pela autora.

Quadro 11: Lista de *hashtags* "#" por dia de coleta

Data da Coleta	*Hashtag* do *Treding Topic* do Dia
01/03/2021	#DAYANEMELLOCAMPEA
02/03/2021	#AForcaDoQuerer
03/03/2021	#ForaArthur
04/03/2021	#DoriaGenocida
05/03/2021	Lexa
06/03/2021	#paredaofalso
07/03/2021	GENOCIDA
08/03/2021	#Bolsonaro2022
09/03/2021	#CarlaNoQuartoFalso
10/03/2021	Lulaland
11/03/2021	#BBB21Carla
12/03/2021	#EuSouOExercitoDoBrasil
13/03/2021	#BolsonaroGenocidaSim
14/03/2021	#CATBBB
15/03/2021	#PLdoGasOriginal
16/03/2021	#FelipeNetoNaCadeia
17/03/2021	#ConteComigoBolsonaro
18/03/2021	#ForaCarla
19/03/2021	#PiranhxsLGBTQ
20/03/2021	#AmorDeMae
21/03/2021	Thelma
22/03/2021	Capitu
23/03/2021	Xuxa
24/03/2021	#NinguemMexeComErnesto
25/03/2021	#ForaJuliete
26/03/2021	Butanvac
27/03/2021	Danilo
28/03/2021	Camilla
29/03/2021	Domenico
30/03/2021	Lurdes
31/03/2021	#OBrasilAmaBolsonaro

Fonte: Elaborada pela autora.

Quadro 12: Lista de candidatos a “novas” Expressões Idiomáticas coletados no mês de março/2021.

Expressões Idiomáticas
Colocar fogo no parquinho
Dar/ter close errado
Dar biscoito
Dar PT
Estar na Disney
Estar/ficar pistola
Estar/ser pago
Fingir demência
Jantar cedo
Passar pano

Fonte: Elaborada pela autora.