



Programa de
Pós-Graduação em
Linguística

ASPECTOS LINGUÍSTICOS NA DESCRIÇÃO DE NOTÍCIAS SATÍRICAS DO
PORTUGUÊS DO BRASIL: UMA PROPOSTA TIPOLOGICA

SÃO CARLOS
2022



Universidade Federal de São Carlos

Gabriela Wick-Pedro

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

Aspectos linguísticos na descrição de notícias satíricas do português do Brasil:
uma proposta tipológica

Gabriela Wick-Pedro
Bolsista: CAPES

Tese apresentada ao Programa de Pós-Graduação em Linguística do Centro de Educação e Ciências Humanas da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutora em Linguística.

Orientador: Prof. Dr. Oto Araújo Vale

São Carlos
2022



UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Gabriela Wick-Pedro, realizada em 27/10/2022.

Comissão Julgadora:

Prof. Dr. Oto Araujo Vale (UFSCar)

Profa. Dra. Cláudia Dias de Barros (IFSP)

Prof. Dr. Dirceu Cleber Conde (UFSCar)

Profa. Dra. Helena de Medeiros Caseli (UFSCar)

Prof. Dr. Thiago Alexandre Salgueiro Pardo (USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Linguística.

Wick-Pedro, Gabriela

Aspectos linguísticos na descrição de notícias satíricas
do português do Brasil: uma proposta tipológica /
Gabriela Wick-Pedro -- 2022.
161f.

Tese de Doutorado - Universidade Federal de São Carlos,
campus São Carlos, São Carlos
Orientador (a): Oto Araújo Vale
Banca Examinadora: Oto Araujo Vale, Cláudia Dias de
Barros, Dirceu Cleber Conde, Helena de Medeiros
Caseli, Thiago Alexandre Salgueiro Pardo
Bibliografia

1. Notícia Satírica. 2. Sátira. 3. Corpus. I. Wick-Pedro,
Gabriela. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática
(SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Ronildo Santos Prado - CRB/8 7325

Às vítimas da desinformação e do negacionismo que foram levadas precocemente de suas famílias pela COVID-19.

AGRADECIMENTOS

Esta não é uma pesquisa apenas minha, pois todo o conhecimento não se faz sozinho. Se faz na sala de aula, tomando café, no laboratório, nos congressos, nos corredores da Universidade, nas conversas informais com os colegas pesquisadores e – até mesmo – na mesa do bar. Contudo, uma parte desta tese foi construída durante uma pandemia, quando de repente nos isolamos de toda a comunidade, transformando o desenvolvimento deste trabalho em um processo solitário. Então todas as pessoas a quem agradeço foram fundamentais nesse momento tão difícil e me ajudaram muito, sobretudo a manter a sanidade – ou pelo menos um pouquinho dela. Começo agradecendo a minha família, por todo o carinho depositado por toda a vida, a paciência, ao carinho e aos abraços quando chorava pelas dores da vida adulta e pelas dificuldades enfrentadas na jornada acadêmica no Brasil.

Agradeço imensamente aos meus pais, Hélio e Silvana, por serem meu porto seguro, pela paciência, pelos momentos de alegria e dor que foram divididos, pela profunda confiança e pela segurança para seguir o meu caminho que sempre encontrei em vocês. Um agradecimento especial à minha mãe, pois sem o seu sorriso e suas palavras, teria dificilmente chegado até aqui. Agradeço também ao meu irmão, Heitor, por me mostrar que podemos ser leves nos momentos mais difíceis. Ao meu tio, Silvio, por sempre vibrar com minhas conquistas. Às minhas avós, Clarice e Dulce (*in memoriam*), por me ensinarem e me construírem, cada uma do seu jeito, como uma mulher forte.

Muito obrigada a todos meus amigos que estiveram comigo durante esse caminho. Em especial, ao Matheus Hebling por toda uma vida de amizade, pelos ‘help’ no inglês e pelas risadas proporcionadas (às vezes fora de hora), à Domila e ao André Luiz, pelos melhores momentos pré-pandemia e pelo carinho de sempre.

Agradeço ao Prof. Dr. Oto Araújo Vale, meu orientador há mais de dez anos, que sempre depositou muita confiança em mim, até mesmo quando eu acreditava não ser capaz de seguir na carreira acadêmica e pelo acolhimento nas horas mais difíceis. Apesar da cara de bravo, sempre esteve disposto a me ouvir e me aconselhar (mesmo com alguns puxões de orelha). Obrigada pela orientação, paciência e, sobretudo, por ter me dado excelentes oportunidades no meu caminho acadêmico.

Às Professoras Doutoras Helena Medeiros Caseli e Cláudia Dias de Barros e aos Professores Doutores Thiago Alexandre Salgueiro Pardo e Dirceu Cleber Conde pela disponibilidade e pelo aceite em fazer parte da banca examinadora desta tese. Destino também meus agradecimentos à Profa. Dra. Claudia Freitas pelas importantes considerações e contribuições durante minha qualificação.

Ao Programa de Pós-Graduação em Linguística da UFSCar, pelo suporte institucional, sobretudo à Vanessa, não só por resolver atenciosamente os problemas de cada aluno, como também pela preocupação e vibrar pelas nossas conquistas e cada um de

nós. Ao Núcleo Interinstitucional de Linguística Computacional (NILC/USP) pelas ajudas profissionais e novas perspectivas linguístico-computacionais. Agradeço também aos colegas do grupo LeGOS (Léxico-Gramática, Opinião e Sentimentos) pela ajuda, discussões no processo desta pesquisa e pelas conversas e experiências trocadas. Um agradecimento especial ao amigo Roney, por toda a colaboração computacional que foi essencial para a construção deste trabalho e pela companhia de viagem. Agradeço também ao Isaac por sempre me socorrer com questões linguísticas, matemáticas, estatísticas e computacionais.

Agradeço ao Prof. Dr. Jackson Souza e à Profa. Dra. Roana Rodrigues, pelas colaborações e discursões fundamentais no nosso grupo de pesquisa. Um imenso agradecimento à Profa. Dra. Carolina Scarton, minha supervisora de estágio na Universidade de Sheffield, pela oportunidade não apenas de conhecer novas experiências acadêmicas em outro país, mas pela atenção, pelo acolhimento e por todo suporte nesse período.

Por fim, agradeço, também a CAPES, pelo apoio financeiro concedido durante todo o período do doutorado. No quadro de desmanche da educação brasileira e a descredibilização às universidades públicas, tive muita sorte por ainda ter minha pesquisa financiada.

Este Trabalho foi executado no Centro de Inteligência Artificial (C4AI-USP) com apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (Processo FAPESP 2019/07665-4) e da IBM Corporation.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 88882.426864/2019-01.

“Tenho duas armas para lutar contra o desespero, a tristeza e até a morte: o riso a cavalo e o galope do sonho. É com isso que enfrento essa e fascinante tarefa de viver.”

(Ariano Suassuna)

RESUMO

A presença de conteúdo enganoso (do inglês, *deception*) na web e em aplicativos de mensagens tem se mostrado um grande problema contemporâneo. Esse contexto gerou algumas iniciativas na Linguística e na Computação para caracterizar linguisticamente textos relacionados e detectar automaticamente sua ocorrência. De acordo com (RUBIN; CHEN; CONROY, 2015), existem três tipos tradicionais de conteúdo enganoso: i) notícias fabricadas: produzidas pelo que é chamado de imprensa marrom ou tabloides; ii) boatos: notícias disfarçadas para enganar o público e podem ser divulgadas por descuido pelas agências de notícias tradicionais e iii) notícias satíricas: notícias parecidas com as notícias reais, porém, criadas para fins de humor. Teoricamente, de acordo com Simpson (2003), a sátira pode ser definida, a partir de uma tríade, como uma prática discursiva que estabelece e resulta uma incongruência irônica entre um alvo satírico, um autor satírico e um público satírico e tem como propósito criticar ou zombar do alvo satírico. Assim, se não reconhecidas como um conteúdo de humor, as notícias satíricas podem criar dificuldades de entendimento e falsas crenças nas mentes de leitores mais desatentos. Detectar uma notícia satírica automaticamente, portanto, mostra-se relevante no viés linguístico-computacional, principalmente somado à deficiência de trabalhos na literatura que consideram a análise computacional da sátira e a inexistência para a Língua Portuguesa. Relata-se aqui a construção de um *corpus* de notícias satíricas e seu paralelo de notícias verdadeiras para português brasileiro. O *corpus* é composto por um *subcorpus* de 150 notícias satíricas (22.963 palavras e 1.212 sentenças) extraídas do site Sensacionalista e outro *subcorpus* de 150 notícias verdadeiras (107.133 palavras e 5.721 sentenças) extraídas de diversos portais on-line de notícias e são correspondentes aos artigos satíricos. O *corpus* total contabiliza 130 mil palavras e 6.900 sentenças. Além disso, este trabalho se propõe a analisar e descrever os aspectos morfosintáticos, a diferença das ocorrências verbais das notícias satíricas, bem como as principais características lexicais encontradas nos artigos satíricos e verdadeiros. Para a realização desta tarefa, o *corpus* foi anotado automaticamente pelo *parser* PALAVRAS (BICK, 2000). Também foram utilizadas as ferramentas NILC-Matrix (LEAL, 2021) para medir a complexidade textual nos textos e o LIWC (PENNEBAKER et al., 2015), que avalia componentes emocionais, cognitivos e estruturais de um determinado texto, baseia-se na utilização de um dicionário contendo classificação de palavras em categorias. Finalmente, espera-se contribuir na descrição linguística de notícias satíricas e criar por meio dos resultados obtidos nesta pesquisa, bases para futuros trabalhos do Processamento de Língua Natural (PLN) focados na identificação automática de conteúdo enganoso para o português do Brasil.

Palavras-chave: Notícia Satírica. Sátira. Notícia Falsa. Pistas Linguísticas. Corpus.

ABSTRACT

The presence of deception on the web and in messaging applications has been a major contemporary problem. This context generated some initiatives in Linguistics and Computing to linguistically characterize related texts and automatically detect their occurrence. According to (RUBIN; CHEN; CONROY, 2015), there are three traditional types of misleading content: i) fabricated news: produced by what is called the brown press or tabloids; ii) rumors: news disguised to deceive the public and can be released by carelessness by traditional news agencies and iii) satirical news: news similar to real news, however, created for humor purposes. Theoretically, according to Simpson (2003), satire can be defined, based on a triad, as a discursive practice that establishes and results in an ironic incongruity between a satirical target, a satirical author and a satirical audience, and whose purpose is to criticize or mock the satirical target. Thus, if not recognized as humorous content, satirical news can create difficulties in understanding and false beliefs in the minds of more inattentive readers. Automatically detecting satirical news, therefore, proves to be relevant in the linguistic-computational bias, mainly added to the deficiency of works in the literature that consider the computational analysis of satire and the inexistence for the Portuguese language. The construction of a *corpus* of satirical news and its parallel of true news for Brazilian Portuguese is reported here. The *corpus* is composed of a *subcorpus* of 150 satirical news (22,963 words and 1,212 sentences) extracted from the Sensationalista website and another *subcorpus* of 150 real news (107,133 words and 5,721 sentences) extracted from several online news portals and corresponding to the articles satirical. The total *corpus* counts 130 thousand words and 6,900 sentences. Furthermore, this work proposes to analyze and describe the morphosyntactic aspects, the difference between the verbal occurrences of satirical news, as well as the main lexical characteristics found in satirical and true articles. To perform this task, the *corpus* was automatically annotated by the PALAVRAS parser (BICK, 2000). The NILC-Matrix tools (LEAL, 2021) were also used to measure the textual complexity in texts and the LIWC (PENNEBAKER et al., 2015), which evaluates emotional, cognitive and structural components of a given text, is based on the use of a dictionary containing sorting words into categories. Finally, it is expected to contribute to the linguistic description of satirical news and to create, through the results obtained in this research, bases for future Natural Language Processing (NLP) works focused on the automatic identification of misleading content for Brazilian Portuguese.

Keywords: Satirical News. Satire. Fake news. Linguistic Clues. Corpus.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de notícia do site norte-americano The Onion	19
Figura 2 – Exemplo de manchete do Sensacionalista nas redes sociais	20
Figura 3 – Exemplo de manchete de notícia verdadeira	21
Figura 4 – Exemplo de manchete de notícia satírica e não satírica	22
Figura 5 – Exemplo de sátira contextualmente dependente	35
Figura 6 – Estrutura triádica da sátira como prática discursiva	36
Figura 7 – Exemplo de notícia real	39
Figura 8 – Exemplo de notícia satírica	40
Figura 9 – Busca do termo fake news no Brasil entre 01/01/2011 a 31/12/2021	47
Figura 10 – Ecossistema da desinformação	50
Figura 11 – Desordem da desinformação	50
Figura 12 – Comparação da importância dos quatro conjuntos de características no nível do parágrafo e no nível do documento	60
Figura 13 – Cabeçalho do site do Sensacionalista	73
Figura 14 – Exemplo do atual site do Sensacionalista	74
Figura 15 – Processo de construção do <i>subcorpus</i> de análise	76
Figura 16 – Processo de extração de características pelo NLTK	77
Figura 17 – Disposição de sentenças no <i>subcorpus</i>	80
Figura 18 – Modelo de anotação	84
Figura 19 – Exemplo do formulário de anotação	85
Figura 20 – Exemplos da saída de uma sentença anotada pelo PALAVRAS	88
Figura 21 – Exemplo de hierarquia <i>psychological process</i> do LIWC	92
Figura 22 – Frequência dos atributos linguísticos	102
Figura 23 – Disposição das características linguísticas no <i>subcorpus</i> satírico	103
Figura 24 – Classes gramaticais extraídas pelo <i>parser</i> PALAVRAS (BICK, 2000)	104
Figura 25 – Estatística de leitura de o <i>corpus</i> conforme o Índice Flesch Brasileiro	109
Figura 26 – Porcentagem de interpretação implícita e explícita	114
Figura 27 – Concordância entre os anotadores	118
Figura 28 – Tipologia de características linguísticas em notícias satíricas	141

LISTA DE QUADROS

Quadro 1 – Modelo de notícias satíricas	41
Quadro 2 – Facticidade e intenção do texto	52
Quadro 3 – Principais características dos trabalhos relacionados	71
Quadro 4 – Descrição dos assuntos abordados nas categorias	75
Quadro 5 – Sátira implícita e explícita	81
Quadro 6 – Categoria das classes gramaticais do PALAVRAS	87
Quadro 7 – Características linguísticas das notícias satíricas	101
Quadro 8 – Leiturabilidade do <i>corpus</i> de acordo com o Índice Flesch Brasileiro . .	109
Quadro 9 – Valores de interpretação do coeficiente <i>kappa</i>	117
Quadro 10 – Exemplos de marcações de sátira nas sentenças	121

LISTA DE TABELAS

Tabela 1 – Características do <i>corpus</i>	75
Tabela 2 – Características do <i>subcorpus</i> (NLTK)	76
Tabela 3 – Características do <i>subcorpus</i> (NILC-Metrix)	78
Tabela 4 – Frequência das características linguísticas	103
Tabela 5 – Relação de tempo verbal entre notícias reais e satíricas	105
Tabela 6 – Relação da pessoa verbal entre notícias reais e satíricas	106
Tabela 7 – Relações psicolinguísticas a partir dos dados extraídos do LIWC	107
Tabela 8 – Frequência de emoções das notícias com base no dicionário do LIWC	108
Tabela 9 – Principais resultados extraídos do NILC-Metrix	110
Tabela 10 – Total de sentenças classificadas como implícita e explícita	114
Tabela 11 – Interpretação implícita e explícita das sentenças das notícias satíricas	115
Tabela 12 – Categorias e anotadores para cálculo do coeficiente <i>kappa</i>	116
Tabela 13 – Concordância da compreensão da sátira entre anotadores	120

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
LIWC	Linguistic Inquiry and Word Count
PLN	Processamento de Língua Natural
PoS	Part-of-Speech
IAA	Inter-Annotator Agreement
ex.s	Exemplos das notícias satíricas
ex.r	Exemplos das notícias reais
Pej	Palavras pejorativas
Vul	Vulgarismos
Aut	Autoria
JPa	Jogo de palavras
Emo	Emoticons/emojis
Rep	Repetição
TSuj	Troca do sujeito
EId	Expressões idiomáticas
FDC	Palavras fora do contexto de domínio
Amb	Ambiguidade
QExp	Quebra de expectativa
Euf	Eufemismo
Met	Metáfora
Exa	Exagero
Pers	Personificação

SUMÁRIO

1	Introdução	18
1.1	Objetivos e hipóteses	24
1.2	Estrutura do texto	25
2	Fundamentação Teórica	27
2.1	Breve introdução da sátira: origem e características	27
2.2	A sátira enquanto objeto linguístico	31
2.2.1	Kreuz e Roberts (1993)	32
2.2.2	Simpson (2003)	35
2.3	A sátira e as notícias satíricas	38
2.4	As notícias satíricas como conteúdo enganoso	44
3	Trabalhos Relacionados	53
3.1	Métodos de PLN voltados às notícias satíricas	53
3.2	As características linguísticas de notícias satíricas	56
4	Métodos e Materiais	72
4.1	<i>Corpus</i> da pesquisa	73
4.1.1	Seleção do <i>corpus</i>	73
4.1.2	Criação do SatiriCorpus.Br	74
4.1.3	Construção do <i>subcorpus</i>	75
4.2	Aspectos linguísticos na descrição das notícias satíricas	79
4.2.1	Classificação das sentenças do <i>subcorpus</i>	79
4.2.2	Classificação da interpretação satírica: explícita e implícita	81
4.2.2.1	Anotação de sentenças explícitas e implícitas	83
4.3	Ferramentas linguístico-computacionais	85
4.3.1	Etiquetagem morfosintática pelo PALAVRAS	85
4.3.2	Medição da inteligibilidade textual pelo NILC-Matrix	88
4.3.2.1	Índice Flesch	89
4.3.2.2	Coh-Matrix	90
4.3.2.3	Coh-Matrix para o português	90
4.3.3	LIWC	91
5	Análise Linguístico-computacional de Notícias Satíricas	95
5.1	Análise humana na descrição das notícias satíricas	95
5.1.1	Características lexicais	95
5.1.2	Característica sintática	97
5.1.3	Características semânticas	98
5.1.4	Características estilísticas	99
5.1.5	Resultados da análise das características linguísticas	101
5.2	Resultados obtidos pelas ferramentas de PLN	104

5.2.1	Características extraídas do PALAVRAS	104
5.2.2	Características extraídas pelo LIWC	106
5.2.3	Características extraídas pelo NILC-Metrix	108
5.2.3.1	Análise Índice Flesch	108
5.2.3.2	Análise NILC-Metrix	109
5.3	Análise das interpretações das sentenças satíricas	113
5.3.1	Anotação de manchetes implícitas e explícitas	115
6	Tipologia de pistas linguísticas em notícias satíricas	123
6.1	Descrição da tipologia de sinais linguísticos em notícias satíricas	123
6.2	Características descritivas	123
6.3	Características morfosintáticas	124
6.4	Características linguísticas	129
6.5	Características psicolinguísticas (LIWC)	132
6.6	Características de inteligibilidade textual (NILC-Metrix)	137
7	Conclusão	142
	Referências	146
	Apêndices	156
	APÊNDICE A Diretrizes de Anotação	157
	APÊNDICE B Tipologia	159

1 INTRODUÇÃO

Nas últimas décadas, a utilização de mídias sociais tem mudado a estrutura de consumo e reprodução de notícias veiculadas na rede. Desse modo, um conteúdo jornalístico pode ser transmitido de uma pessoa para a outra sem uma fiscalização ou uma filtragem editorial (ALLCOTT; GENTZKOW, 2017). Ao passo que o novo modo de compartilhamento de informações possibilita um maior alcance de usuários, também facilita a propagação de desinformação, que abrange conteúdos de naturezas muito diversas, como informações em que não há a certeza da sua fonte ou informações fabricadas para imitar o conteúdo dos grandes meios de comunicação. Recentes pesquisas acadêmicas procuram estudar a desinformação (WARDLE; DERAKHSHAN, 2018), investigar o comportamento (SINTRA, 2019) e o perfil dos usuários que compartilham e produzem esse tipo de notícia e como elas se espalham pela rede (PENNYCOOK et al., 2021).

Em 2017, o Dicionário Collins elegeu o termo *Fake News* a palavra do ano, após se popularizar com as eleições presidenciais nos EUA no ano anterior, definindo-a como “informações falsas, muitas vezes sensacionalistas, divulgadas sob o disfarce de notícias” – tradução nossa¹. Popularmente, consideram-se boatos, rumores e outros tipos de informações enganosas (com intenção ou não de enganar) como notícias falsas, incluindo notícias satíricas. Diante desse cenário, a heterogeneidade do termo e as diversas definições para o conceito de notícias falsas transpassa as limitações conceituais, pois uma notícia pode ser projetada intencionalmente para enganar o leitor, ser criada para atrair cliques e obter lucro ou ser uma notícia satírica com o propósito de entreter o público (RUBIN; CHEN; CONROY, 2015; WARDLE; DERAKHSHAN, 2017; TANDOC; LIM; LING, 2018).

O objetivo das notícias satíricas não é enganar as pessoas, como acontece com as notícias falsas, mas elas podem criar uma falsa crença de verdade na mente dos leitores mais desatentos ou sem uma fundamentação contextual e cultural (RUBIN et al., 2016). De acordo com Rubin et al. (2016), isso acontece porque as falsidades presentes nesse tipo de conteúdo são intencionalmente utilizadas e pedem para serem desvendadas. Entretanto, como cada leitor possui um conhecimento de mundo, algumas relações entre as informações verdadeiras e falsas não são realizadas, e, desse modo, a notícia falsa é disseminada. Além disso, as notícias satíricas são consumidas como uma forma de entretenimento aos leitores por meio do absurdo, do exagero ou do ridículo e, em simultâneo, utiliza-se do humor para informar e censurar a sociedade moderna.

No ano de 2014, por meio da etiqueta “*satire*”, o Facebook apresentou uma forma de marcar as notícias satíricas do *The Onion*² que circulavam na *timeline* de seus usuários. Segundo a BBC³, isso foi realizado porque a rede social estava recebendo o feedback

¹ “false, often sensational, information disseminated under the guise of news reporting”. Disponível em: <<http://collinsdictionary.com/dictionary/english/fake-news>>. Acesso em: 08 de dez. 2021.

² Disponível em: <<https://www.theonion.com/>>.

³ Disponível em: <<https://www.bbc.com/news/technology-28834682>>.

de alguns usuários que queriam um modo mais claro para discernir artigos satíricos de outros nessa seção. O site *The Onion* é um jornal satírico norte-americano e é, atualmente, reconhecido como um dos maiores sites de humor dos Estados Unidos. Como na maioria dos noticiários satíricos, o site se baseia em eventos reais para criar paródias de notícias factuais e se apresenta como artigos de grandes meios de comunicação, como mostra a Figura 1.

Figura 1 – Exemplo de notícia do site norte-americano The Onion



Fonte: Página do The Onion.⁴

A sátira pode ser um modo cômico de criticar um alvo satírico, aparecendo em diferentes mídias, como charges, programas de televisão e noticiário (LAMARRE; LANDREVILLE; BEAM, 2009; SINGH, 2012). Para a Literatura, a sátira é a representação literária de um estilo escrito em verso ou prosa que tem como objetivo a crítica às instituições, à sociedade e aos hábitos culturais de um povo. A sátira é considerada um gênero literário focado na crítica de um determinado tema, utilizando-se da ironia, do sarcasmo e da paródia para apontar falhas morais, políticas e sociais (KREUZ; ROBERTS, 1993; SIMPSON, 2003; ATTARDO, 2014).

A Figura 2 mostra um exemplo de uma manchete compartilhada pelo Sensacionalista em seu Twitter. É importante ressaltar que a notícia publicada pelo Sensacionalista faz referência ao quadro de saúde de Jair Bolsonaro. Nesse período, ele havia sido diagnosticado com obstrução intestinal⁵, quando há o bloqueio parcial ou completo da passagem das fezes pelo intestino. A partir deste acontecimento, o jornal satírico joga com a ambiguidade da expressão “*fazer merda*” para criar o efeito cômico, pois 1) no sentido de ‘defecar’, não

⁴ Disponível em: <<https://www.theonion.com/>>. Acesso em: 11 de jan. 2022.

⁵ Disponível em: <<http://glo.bo/3EzrDuM>>. Acesso em: 15 de ago. 2022.

seria usual, mas poderia ser interpretável e 2) no sentido ‘*fazer besteiras*’ seria usual, e interpretável.

Figura 2 – Exemplo de manchete do Sensacionalista nas redes sociais

Sensacionalista

Brasil se surpreende ao saber que Bolsonaro está com dificuldade para fazer merda



Fonte: Redes sociais do Sensacionalista.⁶

Em particular, muitas frentes vêm sendo exploradas pelo Processamento de Língua Natural (PLN) (WANG, 2017; PÉREZ-ROSAS et al., 2018; MONTEIRO et al., 2018; BHATT et al., 2021) com o propósito de criar recursos capazes de detectar automaticamente notícias falsas. Além disso, sabe-se que a manipulação computacional da linguagem figurada é atualmente uma das tarefas mais desafiadoras do Processamento de Língua Natural (PLN), dado que esse tipo de linguagem é muitas vezes caracterizado por dispositivos linguísticos como, ironia, sarcasmo, metáforas e humor. O seu significado ultrapassa o significado literal, sendo frequentemente difícil a compreensão até mesmo para os seres humanos⁷. A linguagem figurada exige, muitas vezes, um conhecimento de mundo e a familiaridade com o contexto cultural, informações estas que a máquina não pode acessar facilmente. Além disso, a linguagem figurada está em modificação constantemente devido a

⁶ Disponível em: <<https://www.facebook.com/sensacionalista/photos/d41d8cd9/4408345382541628/>>. Acesso em: 11 de jan. 2022.

⁷ É importante ressaltar que pessoas dentro do espectro autista (neuroatípicas) têm dificuldades para interpretar a linguagem verbal e não verbal, como gestos, tom de voz, expressões faciais, sarcasmo e/ou ironia.

mudanças no vocabulário e na própria linguagem, dificultando o treinamento de algoritmos de aprendizado de máquina.

Diante disso, a capacidade de lidar com a linguagem figurada – e mais especificamente com a sátira – é elementar para construir abordagens linguísticas para o reconhecimento de conteúdo enganoso. Tais abordagens linguísticas da sátira são fundamentais para uma interação humano-máquina, melhorando a maneira como os computadores interpretam e respondem a níveis mais abstratos da língua natural, como o humor, a ironia ou a metáfora. Consequentemente, a possibilidade de detectar e tratar adequadamente a sátira seria benéfica para vários campos nos quais uma compreensão profunda dos traços figurados da língua é essencial.

Um ponto relevante para a manipulação de um conteúdo satírico é como os veículos de notícias satíricas funcionam como uma “ressignificação” dos noticiários factuais, tendo em vista que a sátira pode ser usada para criticar e transmitir um sentido subjetivo ao leitor. Logo, encontrar características que diferem estes conteúdos é um ponto de partida para a descrição de uma notícia satírica. Desse modo, um dos possíveis motivos que dificultam a distinção entre conteúdos reais e satíricos é a recorrência de acontecimentos estranhamente verdadeiros noticiados que despertam a curiosidade do público, levando o leitor a se questionar sobre a veracidade daquele ocorrido.

Nesse contexto, a Figura 3 ilustra bem a dificuldade, muitas vezes, em reconhecer um fato de um conteúdo satírico. A notícia refere-se a justificativa do ex-deputado Daniel Silveira à Polícia Federal por violar 36 vezes as regras do uso de tornozeleira eletrônica. Aqui, não se questiona apenas a veracidade do “cachorro roer o carregador da tornozeleira eletrônica”, mas também a informação circular em um veículo tradicional de comunicação.

Figura 3 – Exemplo de manchete de notícia verdadeira

Daniel Silveira diz em depoimento à PF que cachorro roeu carregador da tornozeleira eletrônica

No mês passado, deputado voltou para a cadeia, em Batalhão da PM, no Rio, por desrespeitar uso da tornozeleira quando estava em prisão domiciliar. Nesta quinta, ele depôs à PF.

Fonte: Página do G1.⁸

Desse modo, ao considerar que o absurdo é uma das principais características da sátira, é possível observar uma problemática presente na Figura 4 – que apresenta duas notícias referentes à crise diplomática entre Jair Bolsonaro, atual presidente do Brasil, e Emmanuel Macron, atual presidente da França. Na época, os dois presidentes trocaram

⁸ Disponível em: <bit.ly/3Xt6qJM>. Acesso em: 11 de jan. 2022.

críticas nas redes sociais após Macron questionar Bolsonaro sobre as grandes queimadas na Floresta Amazônica naquele ano.

Figura 4 – Exemplo de manchete de notícia satírica e não satírica



Fonte: Redes sociais do Sensacionalista.⁹

As duas manchetes apresentadas¹⁰ na Figura 4 são:

- (1) Após fala de Macron, Bolsonaro proíbe venda de pão francês em todo território nacional. [ex.s]
- (2) Bolsonaro diz que vai adotar caneta Compactor porque “Bic é francesa”. [ex.r]

Nota-se nos exemplos acima que existe uma semelhança entre as duas sentenças – há uma contrariedade de Bolsonaro em relação a produtos franceses. No entanto, o exemplo (1) é uma notícia satírica, criada pelo site Sensacionalista e o exemplo (2) é uma notícia factual, veiculada pelo portal de notícias *Correio Braziliense*¹¹. A partir dos exemplos, percebe-se que o momento político em que mundo e, principalmente, o Brasil se encontra pode dificultar ainda mais a tarefa de detecção de notícias satíricas.

Como já foi observado, uma notícia satírica pode ser confundida com uma verídica por leitores que não alcançaram o efeito de humor e ironia que os textos de sátira buscam atingir. Muitas vezes, essa dificuldade de distinção entre o conteúdo factual e o conteúdo

⁹ Disponível em: <<https://twitter.com/sensacionalista/status/1167261092676014085>>. Acesso em: 11 de jan. 2022.

¹⁰ Os exemplos apresentados nesta tese foram, em sua maioria, retirados do *subcorpus* descrito na Subseção 4.1.3. Logo, para a distinção entre as notícias reais e satíricas, serão utilizadas as notações “[ex.r]”, para os exemplos das notícias reais e “[ex.s]”, para os exemplos das notícias satíricas.

¹¹ Disponível em: <<https://bit.ly/3qexYTZ>>. Acesso em: 15 de nov. 2021.

humorístico ocorre porque dentro de uma notícia satírica pode haver informações e/ou acontecimentos reais inseridos ao enunciado satírico. Esta sobreposição de informações reais e fictícias pode confundir o leitor.

Na notícia representada pelo exemplo (3), o trecho da manchete “*Bolsonaro tem 30% de ruim ou péssimo*” é uma informação verdadeira¹², assim como toda a primeira sentença da notícia. Já os trechos sarcásticos que marcam a comparação de Bolsonaro com “*síndico do prédio*” e com o “*barulhinho da broca do dentista*” deixam evidente ao leitor de que a notícia não é verdadeira.

- (3) Bolsonaro tem 30% de ruim ou péssimo e 70% fugiram só de ouvir seu nome
Jair Bolsonaro registrou a maior reprovação de um presidente desde à redemocratização. Com 30% de ruim ou péssimo, ele superou até mesmo Collor, segundo pesquisa da Folha de São Paulo divulgada hoje. Bolsonaro só superou o síndico do prédio como figura mais indesejada. Até mesmo o barulhinho da broca do dentista é mais querido do que ele. De acordo com o estudo, se um atendente de telemarketing que desliga assim que você atende se candidatasse, teria mais chances que Bolsonaro. [ex.s]

No entanto, em (4), a sátira presente na notícia é muito mais sutil em relação ao exemplo anterior. Primeiramente, as informações de que Jair Bolsonaro dispensou mais de 8 mil médicos do programa Mais Médicos¹³, de que 24 milhões de brasileiros ficaram sem atendimento médico¹⁴, de que Bolsonaro havia anunciado novas regras do programa¹⁵ e de que ele criaria uma carreira de estado para os médicos brasileiros¹⁶ são informações verdadeiras. Portanto, nota-se que o efeito de humor ocorre na última sentença com “*se estapear*”, além do jogo de palavras “*Menos Médicos*” em referência ao “*Mais Médicos*” e precisam de um conhecimento prévio do assunto para serem identificadas pelo leitor.

- (4) Bolsonaro lança o Menos Médicos
O presidente eleito, Jair Bolsonaro, anunciou ontem a criação do seu novo programa, o Menos Médicos. Cerca de 8 mil médicos serão dispensados e 24 milhões de brasileiros ficarão sem atendimento. O anúncio aconteceu logo depois de Bolsonaro dizer que vai mudar as regras do atual Mais Médicos. O presidente vai criar carreira de estado para médicos brasileiros. Os profissionais brasileiros devem se estapear disputando vagas com salário altíssimo para atender moradores de cidades remotas e populações ribeirinhas. [ex.s]

Um dos maiores desafios da detecção automática de um conteúdo satírico é a ruptura linguística em vários níveis descritivos da língua – da escolha lexical, estrutura

¹² Disponível em: <<https://bit.ly/3QB9saB>>.

¹³ Disponível em: <<http://glo.bo/3KSE1XU>>.

¹⁴ Disponível em: <<http://glo.bo/3qey8L5>>.

¹⁵ Disponível em: <<https://bit.ly/3qeZIYv>>.

¹⁶ Disponível em: <<https://bit.ly/3KQSTWy>>.

sintática, semântica até a conceitualização. Assim, não é realista buscar uma bala de prata linguístico-computacional para a descrição da sátira, da ironia, do humor ou de qualquer outro recurso figurado da língua. Destarte, uma resposta geral será dificilmente encontrada em apenas uma única técnica ou um único algoritmo. Em vez disso, deve-se identificar aspectos específicos e dispositivos linguísticos suscetíveis a análise computacional, para então, a partir de tratamentos individuais, tentar sintetizar uma solução gradualmente mais ampla.

Apesar de as notícias satíricas terem sido tema de investigações em diversas áreas¹⁷, não se conhece trabalhos focados na análise linguístico-computacional da sátira noticiosa para o português do Brasil. Dessa forma, ao considerar os problemas elencados nesta seção, como a propagação da desinformação nas redes sociais, a classificação das notícias falsas e as particularidades das notícias satíricas, bem como sua dificuldade de compreensão, não só para os humanos, mas principalmente para a máquina, neste estudo, espera-se compreender melhor como ocorre a construção do efeito satírico e quais os possíveis mecanismos linguísticos utilizados como recursos para gerar a sátira e o humor nesses textos, bem como as principais diferenças entre as notícias satíricas e as factuais.

1.1 OBJETIVOS E HIPÓTESES

Este trabalho tem por objetivo descrever e analisar a sátira em notícias com vistas a contribuir para os estudos linguísticos e computacionais do processamento da linguagem figurada, assim como fornecer recursos para trabalhos do PLN com foco na detecção automática de conteúdo enganoso. Com base nisso, propõe-se uma descrição linguística de características capazes de identificar quando uma notícia é satírica ou não a partir de um *corpus*, criado para esta tese, composto por notícias satíricas retiradas automaticamente do site Sensacionalista¹⁸.

A compreensão da sátira está em níveis descritivos da língua mais complexos de processamento automático, como o nível semântico, que busca estudar o significado das palavras, de acordo com o contexto inserido, e pragmático, que estuda a língua em uso, ou seja, a linguagem em seu contexto comunicacional, seja escrita, falada ou sinalizada. Além disso, a descrição da notícia satírica ocorre por meio de dois eventos conflitantes – o evento real e o evento satírico – e esse conflito pode ser marcado textualmente em níveis menos abstratos da língua. Tendo em mente que a compreensão da sátira é inerente ao contexto, entende-se aqui, a necessidade de investigar o comportamento das construções de sentido de uma notícia satírica construída a partir de sinais linguísticos que carregam em sua estrutura elementos de ironia, sarcasmo, sátira e humor.

Especificamente, objetiva-se:

¹⁷ As pesquisas quem abordam as notícias satíricas, em suma, pertencem a áreas da Comunicação, da Linguística Textual ou Análise do Discurso e serão abordadas na Seção 2.3.

¹⁸ Disponível em: <<https://blogs.oglobo.globo.com/sensacionalista>>.

- Criar um *corpus* composto por notícias satíricas;
- Identificar dispositivos linguísticos presentes na estrutura e na superfície do texto que indiquem uma notícia satírica;
- Produzir uma descrição das notícias satíricas, observando as características linguísticas que indiquem uma notícia satírica de uma não satírica;
- Propor uma tipologia para descrição de notícias satíricas que possa servir de subsídio para a detecção automática dessas notícias.

Assim, a partir dos direcionamentos acima, foram levantadas duas hipóteses sobre o fenômeno da sátira nas notícias e sua identificação automática no cenário das notícias falsas:

Hipótese 1 (H1): Uma vez que as características lexicais, sintáticas e semânticas diferem entre análises de notícias satíricas e não satíricas, há sinais linguísticos capazes de identificar linguisticamente uma notícia satírica.

Hipótese 2 (H2): O entendimento do sentido satírico se manifesta por sinais linguísticos de que ora estão presentes na estrutura do texto, ora são dependentes do contexto e do conhecimento extralinguístico.

Para alcançar o objetivo geral e os objetivos específicos que sustentam as hipóteses desta tese, além de considerar a dificuldade da compreensão da sátira apresentada neste capítulo, espera-se responder às seguintes questões:

1. É possível identificar sinais linguísticos que diferem um conteúdo satírico de um não satírico?
2. É possível extrair, a partir de ferramentas de PLN, sinais linguísticos presentes na superfície das notícias que possam ser usados para diferenciar as satíricas das não satíricas?

Desse modo, busca-se não apenas descrever o efeito satírico nas notícias, mas também se espera com esta pesquisa entender melhor sua compreensão e o seu desenvolvimento, levando à descoberta de características e padrões que podem ser compartilhados e confrontados futuramente com projetos semelhantes.

1.2 ESTRUTURA DO TEXTO

A tese está organizada da seguinte forma:

No Capítulo 2 são apresentadas as principais noções e conceitos sobre a sátira. Na primeira parte do capítulo, discutem-se os principais conceitos linguísticos sobre a sátira,

abordando também questões sobre as notícias satíricas e o que elas representam para a Linguística Computacional. Além disso, é mostrada a conceitualização da desinformação. No Capítulo 3 apresentam-se os trabalhos relacionados sobre a descrição e detecção automática de notícias satíricas para outros idiomas, apoiando o desenvolvimento da presente investigação. No Capítulo 4, apresenta-se a metodologia e os materiais utilizados para a construção desta tese. O Capítulo 5 é dedicado à apresentação da análise dos principais aspectos linguísticos presentes em uma notícia satírica. Além disso, são discutidos os resultados preliminares obtidos. No Capítulo 6 são descritas as características linguísticas que compõem a tipologia desenvolvida nesta pesquisa. No Capítulo 7, finaliza-se o texto com as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos da fundamentação teórica que embasam esta tese. Na Seção 2.1 é discutida a origem e as características da sátira enquanto objeto da Literatura. Na Seção 2.2, são introduzidos os principais conceitos que fundamentam a sátira por uma perspectiva linguística, distinguindo-a da paródia e da ironia e apresentando um modelo discursivo para ela. Na Seção 2.3, abordam-se os principais conceitos sobre o que são as notícias satíricas e as suas denominações. Por fim, na Seção 2.4 são detalhados os tipos de conteúdo enganoso e como as notícias satíricas se encaixam nesse tema.

2.1 BREVE INTRODUÇÃO DA SÁTIRA: ORIGEM E CARACTERÍSTICAS

Para compreender o que se entende por sátira atualmente, vê-se necessário um breve embasamento histórico para esclarecer alguns pontos e definições importantes sobre o tema, dado que o objeto norteador deste trabalho incide sobre textos jornalísticos que se utilizam da sátira para criticar eventos reais da atualidade. A relevância dessa discussão se dá pela sátira ser um fenômeno multifacetado, portanto, defini-la não é uma tarefa simples e “propor uma definição universalmente aceitável de sátira é tão provável quanto esgotar o deserto do mar da Odisseia”¹ (ŠSALAMOUN, 2019, p. 30 – tradução nossa).

Etimologicamente o termo sátira tem origem latina, mas as comédias e tragédias gregas também forneceram uma importante referência para a sua consolidação. Com Aristófanes, na peça *Lisístrata*, em 411 a.C., o humor é utilizado para mostrar o quão tola era a guerra entre Atenas e Esparta e, embora o humor já estivesse consolidado na Grécia, foi no período romano que a sátira passou a ser identificada como uma manifestação cômica autônoma. Procedente da forma feminina do adjetivo *satur* (saciado, farto), o substantivo *satura* estaria relacionado ao composto *satura lanx*, que designava um prato repleto de frutos e legumes variados; e também a expressão *per saturam*, um termo político, significando um conjunto de leis que se pretendia votar em conjunto (ULLMAN, 1913). Assumindo um sentido figurado, como termo literário, *satura* significa “mistura, miscelânea” (HIGHET, 1962).

Com o famoso trecho *satira quidem tota nostra est* – a sátira é realmente nossa, Quintiliano (I d.C.) afirma que a sátira é um gênero inteiramente latino. Porém, foi principalmente com Lucílio – e também poetas romanos como Horácio, Pérsio e Juvenal – que se atribuiu à sátira um conceito específico, sendo considerada um gênero literário, no qual as instituições ou pessoas eram criticados e os defeitos da sociedade eram ridicularizados. D’Onofrio (1968) reafirma a origem latina da sátira e a determina:

A sátira, efetivamente, surge da observação dos vícios e das distorções sociais e morais. [...] Numa gama variada de sentimentos,

¹ “[...] proposing a universally acceptable definition of satire is as probable as exhausting the sea god of Odyssey.”

que vai da violência da invectiva até o fino humorístico, **o autor satírico serve-se do ridículo para a finalidade catártica da correção dos costumes**. A sátira, portanto, quer pela sua fonte psicológica (a indignação) quer pelo seu meio expressivo (o ridículo) quer pela sua finalidade (a moralização), não pode ser imitação livresca, porque é a imitação da vida contemporânea ao poeta, o retrato de uma sociedade colhida em sua flagrante atualidade, a descrição de vícios e defeitos peculiares aos homens daquele tempo e daquele lugar. (D'ONOFRIO, 1968, p. 16 – grifo nosso).

D'Onofrio (1968, p. 9) ainda delimita a sátira como “gênero literário que mais retrata a sociedade contemporânea, [pois] lança mão deste cabedal de cultura, utilizando os temas, os motivos e as formas de expressão que estão na moda”. Também partindo da concepção da sátira como um gênero literário, Silva (1997, p. 44) atribui uma dupla constituição da sátira: de um lado, é vista como um gênero moralizante e com o objetivo de melhorar o ser social; de outro, a sátira é concebida para “provocar o riso, a partir da gozação, da ironia, da raiva provocada, tudo o que corroa a aparência dos falsos valores, cultivados sob a máscara da hipocrisia”.

Entretanto, Llera (1999) afirma que dada a multiplicidade de formas em que a sátira se apresenta, não se pode considerá-la estritamente como um gênero literário. Conforme o autor, é praticamente unanimidade entre os teóricos modernos a atribuição da sátira como uma modalidade literária, pois atravessa a estrutura de uma obra ou de um gênero. Nesse contexto, a sátira pode ser considerada um elemento “multiforme” (ROCHA, 2006, p. 13), pois se faz “presente pelos mais diversos veículos, sejam eles artísticos ou não. Hodgart (1969, p. 12) especifica que a sátira se distingue de outros gêneros literários pela forma como aborda o tema a ser tratado, ressaltando que:

O satirista não pinta uma imagem objetiva dos males que descreve, pois o realismo puro seria muito opressivo. Em vez disso, ele geralmente nos oferece uma caricatura da situação, que, ao mesmo tempo, dirige nossa atenção para a realidade e permite escapar dela. Toda boa sátira contém um elemento de ataque agressivo e uma visão fantástica do mundo transformado: é escrita para entretenimento, mas contém comentários contundentes e reveladores sobre os problemas do mundo em que vivemos, oferecendo ‘jardins imaginários com sapos reais neles’. Parece, então, que a sátira é distinguível de outros tipos de literatura por sua abordagem de seu assunto, por uma atitude especial em relação à experiência humana que se reflete em suas convenções artísticas. De fato, é muito difícil distingui-la nitidamente de outras formas literárias em qualquer outra base; primeiro, porque não forma explicitamente um dos “gêneros” tradicionais e, segundo, porque pode assumir uma variedade desconcertante de sub-formas. (HODGART, 1969, p. 12 – tradução nossa)².

² “The satirist does not paint an objective picture of the evils he describes, since pure realism would be too oppressive. Instead he usually offers us a travesty of the situation, which at once directs our attention to actuality and permits an escape from it. All good satire contains an element of aggressive attack and a fantastic vision of the world transformed: it is written for entertainment, but contains sharp and telling comments on the problems of the world in which we live, offering imaginary gardens with real toads in them. It would seem, then, that satire is distinguishable from other kinds of literature by its approach to its subject, by a special attitude to human experience which is reflected in its

Hodgart (1969) aponta que por aparecer em variadas formas, não é possível chegar a um acordo sobre a definição estrita da sátira. No entanto, ela tem o objetivo de expor aspectos negativos com o propósito de corrigi-los através do incômodo causado. O autor ainda destaca que a sátira deve ser reconhecida por sua capacidade de abstração, pela fantasia que a compõe, pelo papel revelador sobre os problemas e questionamentos que a sociedade vive (HODGART, 1969).

Knight (2004) também abre mão da ideia da sátira como gênero literário. O autor afirma que se trata de “um explorador de outros gêneros” e também assume a multiformidade, visto que “a sátira é uma posição mental que precisa adotar um gênero para expressar suas ideias como representação”³. Além disso, para Knight (2004):

Seu elemento característico de ataque é muitas vezes formal: o satirista significa o ataque, mas também pode usar o ataque para implicar mais significado. As complexas manipulações de formas e linguagem da sátira para chegar e apresentar sua representação negativa estabelecem a natureza de sua moldura. (KNIGHT, 2004, p. 4 – tradução nossa)⁴

Assim, ao considerar que a sátira pode se manifestar em diversos gêneros além de versos e poemas⁵, Stinson (2019) evidencia que a sátira é uma forma literária intencional e proposital, e não um gênero; é uma forma tanto humorística quanto crítica por natureza, tendo como base em seu comportamento o ataque a alvos específicos e reais. Na verdade, quando estudada, é praticamente um consenso: todas as concepções ou teorias concordam que a sátira parte da raiva e da indignação; e que através do humor decide corrigir, censurar e ridicularizar as imprudências e desregramentos da sociedade. Uma obra satírica não se limita apenas aos estudos literários, mas assume amplos modos de expressão, indo desde as artes literárias às peças teatrais, desenhos animados, notícias satíricas, filmes e programas de televisão. Para Fowler (1982, p. 110), a “sátira é o modo mais problemático para o taxonomista, pois parece nunca ter correspondido a nenhum tipo. Pode assumir quase qualquer forma, e claramente vem fazendo isso há muito tempo”⁶.

artistic conventions. It is in fact very difficult to distinguish it clearly from other literary forms on any other basis; first, because it does not clearly form one of the traditional genres, and secondly because it may assume a bewildering variety of sub-forms.”

³ “[...] a sátira é uma posição mental que precisa adotar um gênero para expressar suas ideias como representação.”

⁴ “Its characteristic element of attack is often formal: the satirist means the attack but may also use the attack to imply further meaning. Satires complex manipulations of forms and language in order to arrive at and present its negative representation establish the nature of its frame.”

⁵ Na Literatura, são considerados três gêneros literários clássicos: narrativo ou épico, lírico e dramático. Desse modo, a sátira, em algum momento, deixou de estar presente nesses gêneros clássicos, sendo algo empregada também sobre outros gêneros. Por exemplo, a obra Dom Quixote trata-se de uma sátira às antigas novelas de cavalaria, considerada uma das maiores obras da literatura espanhola. No cinema, filmes do grupo Monty Python, como a Vida de Brian, uma sátira da história do Rei Arthur e os Cavaleiros da Távola Redonda, usam o gênero como meio do humor.

⁶ Satire is the most problematic mode to the taxonomist, since it appears never to have corresponded to any one kind. It can take almost any external form, and has clearly been doing so for a very long time.”

Atualmente, a sátira pode ser representada em diversas maneiras através de programas televisivos que fornecem uma reação animada satírica a tópicos atuais, como desenhos animados (*South Park*, *Os Simpsons*, *Family Guy*, etc), telejornais (*Saturday Night Live*, *The Daily Show*) e jornais de notícias tradicionais (*The Onion*, *Sensacionalista*, *Piauí Herald*). No entanto, há ainda uma grande variedade de plataformas na mídia moderna, como peças de teatro, livros, filmes e revistas que se utilizam da sátira para construir seu conteúdo. Ainda que existam também antigos tipos de jornais satíricos, como brasileiro *O Pasquim* e o francês *Le Canard enchaîné*, eles não serão abordados neste trabalho. Para os propósitos desta tese, apenas notícias satíricas atuais do português brasileiro serão utilizadas como objeto de estudo.

Um ponto em comum entre as expressões satíricas retratadas acima é a utilização do humor e do efeito cômico, no qual diferentes elementos da linguagem, cultura e sociedade são de grande importância. Attardo (1994, p. 4) afirma que os linguistas, psicólogos e antropólogos tomaram o humor como uma ampla categoria que abrange qualquer evento ou objeto que provoca riso, diverte ou é sentido como engraçado. Ainda para Attardo (1994), o humor pode ser definido por sua função, que, em sua opinião, é provocar o riso no público.

Conforme assinala Greenberg (2019), a sátira é considerada uma espécie característica de comédia, mas difere-se devido ao seu caráter intencional, crítico ou tendencioso. Bergson (2018) reitera a função do riso como um modo social de correção. Segundo o autor, o humor e a ironia são opostos e ambos são formas de sátira, mas enquanto a ironia é de natureza oratória, o humor tem algo de mais científico. Desse modo, à medida que a ironia se arrasta sob a ideia do bem, e de como as coisas deveriam ser na realidade, o humor desce até ao interior do mal, onde analisa todas as particularidades com indiferença (BERGSON, 2018, p. 92). Por fim, Hansen (2004) complementa que a sátira não deve ser fundamentalmente engraçada, mas que se utiliza de práticas verossimilhantes para que, através do humor, ela possa adquirir uma postura crítica dos acontecimentos.

Outra questão a ser destacada é que algumas manifestações satíricas, como as representadas em telejornais/noticiários, são paródias que fazem alusão à estrutura e conteúdo real veiculados nos meios verídicos. Nas palavras de Hodgart (1969), a paródia é a base de toda sátira literária, o que envolve assumir e dominar o estilo de outro escritor e reproduzi-lo com distorções ridículas. Frequentemente, a sátira acaba sendo confundida com a paródia e uma razão para esta confusão entre paródia e sátira “é o fato de os dois gêneros serem muitas vezes utilizados conjuntamente (HUTCHEON, 1985, p. 62). A autora ainda afirma que:

A sátira usa, frequentes vezes, formas de arte paródicas, quer para fins expositórios, quer para fins agressivos (PAULSON 1967, 5-6), quando aspira à diferenciação textual como veículo. Tanto a sátira como a paródia implicam distanciação crítica e, logo, julgamentos de valor, mas a sátira utiliza geralmente essa distância para fazer uma afirmação

negativa acerca daquilo que é satirizado – “para distorcer, depreciar, ferir” (HIGHET, 1962, 69). Na paródia moderna, no entanto, verificamos não haver um julgamento negativo necessariamente sugerido no contraste irônico dos textos. A arte paródica desvia de uma norma estética e inclui simultaneamente essa norma em si, como material de fundo. Qualquer ataque real seria autodestrutivo (HUTCHEON, 1985, p. 62).

Greenberg (2019) reitera a paródia como um trabalho que imita outro trabalho de forma humorística ou lúdica, salientando que a paródia representa outros textos e outras obras de arte, enquanto a sátira representa o próprio mundo. Nesse sentido, segundo o autor, “os textos são parte do mundo e, assim, ao parodiar os textos, os escritores muitas vezes satirizam as ideias, valores ou atitudes neles incorporados” (GREENBERG, 2019, p. 33 – tradução nossa)⁷. Alinhando-se a Hutcheon (1985), para Greenberg (2019, p. 33 – tradução nossa), “as categorias se sobrepõem: a paródia é frequentemente citada como uma “técnica” de sátira ou mesmo como um subtipo, e o ato de paródia pode dobrar como um ato de sátira”⁸.

Enfim, ressalta-se a ironia como uma das características mais contundentes da sátira. Toma-se como definição de ironia a sua definição tradicional: a ironia consiste na inversão semântica do sentido, dizendo o contrário do que realmente se diz (DUARTE, 2006; MUECKE, 1995). A sobreposição entre ironia e sátira é tão evidente que em certas circunstâncias funcionam como sinônimos (GREENBERG, 2019) e tanto a sátira quanto a ironia são caminhos indiretos para a verdade (WEISGERBER, 1973). Um exemplo do uso da ironia em uma obra satírica é a ironia⁹ presente em “(As) Viagens de Gulliver” de Jonathan Swift. Quando Gulliver visita o País dos Houyhnhnms e encontra uma raça de cavalos que são racionais, ele mantém os humanos como sua força bruta. É essa inversão de papéis que Swift utiliza para ironizar a sociedade em que vivia.

Por fim, considerando que a sátira transpassa a classificação como um gênero literário, sendo mais um modo de se expressar sobre as polêmicas, desgostos e críticas a um evento da sociedade, a próxima seção discute como os recursos da linguagem (tratado por Kreuz e Roberts (1993), como ironia e paródia) são necessários para a construção da sátira e suas diferenciações. Também é apresentado o modelo discursivo de Simpson (2003) para a sátira.

2.2 A SÁTIRA ENQUANTO OBJETO LINGUÍSTICO

Os noticiários satíricos constituem um interessante e promissor campo de pesquisa em diversas áreas acadêmicas, especificamente, nos estudos da linguagem que buscam um

⁷ “[...] texts are part of the world, and so in parodying texts writers often thereby satirize the ideas, values, or attitudes embodied in them.”

⁸ “[...] the categories overlap: parody is often cited as a technique of satire or even as a subtype, and the act of parody can double as an act of satire.”

⁹ Na obra “(As) Viagens de Gulliver”, Jonathan Swift se utiliza da ironia situacional, i.e., ocorre quando o resultado real de uma situação é totalmente diferente do que você esperaria que fosse.

olhar descritivo dos meios linguísticos e estilísticos da construção da sátira. Como já discutido, a definição de sátira é bem elusiva, pois existe uma disparidade em seu entendimento, principalmente, entre aqueles que consideram a sátira como um gênero literário e aqueles que acreditam ser um modo de se manifestar, muitas vezes parodicamente, em outros gêneros e discursos. Compreender a sátira vai além de delimitá-la em gênero literário ou não, reconhecer seus objetivos e suas principais técnicas. É necessário compreender quais são suas características linguísticas e como a sátira se constrói como um elemento textual, um discurso satírico.

Assim, nesta seção são abordadas as principais perspectivas linguísticas sobre a sátira que sustentam este trabalho. Como visto na seção anterior, a sátira é composta de recursos estilísticos como a ironia, a paródia e o humor. Desse modo, entende-se a importância de compreender melhor estes elementos, quais suas diferenças e como eles se constroem na sátira. Na discussão de (KREUZ; ROBERTS, 1993) são apresentadas as discussões e relações entre a sátira, paródia e ironia, salientando que os autores consideram a paródia como gênero, enquanto a ironia é definida como um complexo dispositivo retórico usado por esses gêneros. Finalmente, para a compreensão da sátira, é introduzido o modelo de (SIMPSON, 2003) que analisa o humor, explorando o discurso satírico. Para o autor, a sátira é entendida como uma prática discursiva e não como um gênero do discurso. Esta é a premissa sob a qual esta tese opera.

2.2.1 Kreuz e Roberts (1993)

Em uma visão comum dos mecanismos figurados da língua, para a maioria dos estudiosos da linguagem, não há uma distinção clara em relação aos limites de diferenciação entre paródia e sátira. Nesse sentido, Kreuz e Roberts (1993, p. 98) afirmam que “parte da confusão sobre os conceitos de sátira e paródia resulta do fato delas compartilharem características de outro conceito mal compreendido e frequentemente mal interpretado: a ironia” (tradução minha). Os autores defendem que a ironia – um complexo artifício retórico – pode ser usada tanto pela paródia quanto pela sátira.

Apesar de a ironia não ser o foco desta pesquisa e como já dito anteriormente, ela é considerada, segundo Kreuz e Roberts (1993), como um dispositivo importante para a compreensão e distinção da sátira e da paródia, pois, são frequentemente empregadas nestes gêneros. Ao explicar a ironia, os autores a definem em quatro tipos diferentes. A primeira delas, a *ironia socrática*, corresponde ao método socrático, estratégia pautada na dissimulação de Sócrates, i.e., questionamentos que levavam o interlocutor a uma confusão, apontando-lhe as fraquezas de seus argumentos (SILVA, 1994). Logo, o fingimento é um aspecto importante para esse tipo de ironia (KREUZ; ROBERTS, 1993, p. 98). A segunda, a *ironia dramática* é quando existe uma diferença de consciência entre a personagem de uma obra literária e seu leitor; quando o público possui informações que as personagens não possuem. A terceira é a *ironia do destino*, quando afirmações literais são usadas

para ressaltar uma relação peculiar entre dois eventos. Gibbs (1994) chama a ironia de destino de *ironia situacional* e complementa que ela será o resultado do reconhecimento de estranheza de uma dada situação, havendo discrepância entre o resultado esperado e o resultado real. Por último, a *ironia verbal* é quando o falante diz deliberadamente o oposto do seu significado literal.

Kreuz e Roberts (1993) defendem ainda que tanto a paródia quanto a sátira podem ser descritas pela menção ecoica (SPERBER; WILSON, 1981), pelo fingimento (CLARK; GERRIG, 1984) e pelas representações mentais (JOHNSON-LAIRD, 1983). Assim, na menção ecoica proposta por Sperber e Wilson (1981), o falante ecoa de modo implícito, remetendo um pensamento em que é atribuído a outros pensamentos – sejam eles reais ou não – para expressar sua atitude crítica ou ridicularizada, dando-a como falsa, irrelevante ou pouco informativa. O eco pode ocorrer após o que foi dito, mas também pode estar retomando pensamentos reais ou imaginários. Em uma noção mais ampla, o eco tem suas restrições definidoras. Assim, uma representação acessível não pode ser definida como eco, tendo em vista que ele será lembrado a partir da inacessibilidade à representação na qual haverá uma checagem com a relevância do enunciado.

Em contrapartida, Clark e Gerrig (1984) acreditam que a teoria da menção ecoica não consegue explicar todas as situações irônicas, uma vez que é necessário compartilhar um mundo comum entre os falantes do enunciado irônico e não apenas nomeá-los de forma ecoica. Nesse ponto, enquanto Clark e Gerrig defendem Grice (1975) e suas máximas e que aquele que é irônico está fingindo usar aquele enunciado, os autores também criticam a teoria de Sperber e Wilson (1981), que considera que a ironia como processamento posterior na mente do leitor, uma vez que elas veem a mente como um sistema computacional. Segundo Grice (1975), quando dois indivíduos estão em um diálogo, existem leis implícitas que regem a comunicação. Os interlocutores, mesmo involuntariamente, agem seguindo normas comuns que caracterizam uma cooperação para que a informação possa ser trocada de forma mais homogênea possível. Para ele, estas regras são chamadas de Princípio da Cooperação. Este princípio deve ser seguido em toda produção de um enunciado respeitando quatro máximas:

- **Máxima de Qualidade:** não diga o que você não acredita ser falso; não diga senão aquilo que você possa fornecer evidência adequada;
- **Máxima de Quantidade:** faça com que a sua contribuição seja tão informativa quanto requerido para o propósito corrente da conversação; não faça a sua contribuição mais informativa do que é requerido;
- **Máxima de Relação:** seja relevante e
- **Máxima de Modo:** seja claro (evite ambiguidades, obscuridade de expressão, seja abreviado e ordenado).

Clark e Gerrig (1984) explicam a teoria do fingimento como:

Suponhamos que F está falando com O, o principal ouvinte, e com O', que pode estar presente ou ausente e pode ser real ou imaginário. Ao falar ironicamente, F está fingindo ser F' falando com O'. O que F' está dizendo é, de uma forma ou de outra, claramente uniforme ou imprudente, digno de “um julgamento hostil ou depreciativo ou um sentimento como indignação ou desprezo” (GRICE, 1978, p. 124). Pretende-se que O' em ignorância, não entenda o fingimento e leve F a sério. Mas, pretende-se também que O, como parte do “círculo interno” (nas palavras de Fowler), veja tudo – o fingimento, a imprudência de F', a ignorância de O' e por isso a atitude de F em relação a F', O' e o que F' disse. F' e O' podem ser indivíduos reconhecidos (...) ou pessoas de tipos reconhecidos (como políticos oportunistas) (CLARK; GERRIG, 1984, p. 122 – tradução nossa)¹⁰

A sátira e a paródia também podem ser compreendidas pelas representações mentais (JOHNSON-LAIRD, 1983), modelos de representar o mundo exterior, visto que as pessoas não captam o mundo externo diretamente, elas constroem representações mentais dele. Os modelos mentais, propostos por Johnson-Laird (1983), são blocos de construções cognitivas que podem ser combinados e recombinados conforme necessário.

Com relação à paródia, ela pode ser definida como uma imitação de algo ou alguém com a intenção de ridicularizar e/ou criticar (HOLMAN; HARMON, 1992) *apud* (KREUZ; ROBERTS, 1993) e em oposição à ironia do destino, assim como a sátira, a paródia é um comentário implícito (KREUZ; ROBERTS, 1993). Os autores ainda complementam que:

(...) a sátira e as formas de ironia compartilham muitas das mesmas características, especialmente o uso de fingimento e a necessidade de múltiplas representações mentais. (...) Embora a paródia também exija a construção de múltiplas representações mentais, ela difere da sátira porque se baseia na menção ecoica e não na simulação (KREUZ; ROBERTS, 1993, p. 102 – tradução nossa)¹¹.

Em suma, Kreuz e Roberts (1993) argumentam que tanto a paródia quanto a sátira exigem que o leitor construa múltiplas representações mentais; no entanto, na paródia, o público não precisa ir além dos limites da obra original para considerar as implicações sociais como fazem na sátira. Assim, a paródia mantém o gênero e a interlocução com o texto original, o enredo, os personagens, etc., entende-se, portanto, que a paródia tem um

¹⁰ “Suppose S is speaking to A, the primary addressee, and to A', who may be present or absent, real or imaginary. In speaking ironically, S is pretending to be S' talking to A'. What S' is saying, in one way or another, patently uniformed or injudicious, worthy of a “hostile or derogatory judgment or a feeling such as indignation or contempt” (GRICE, 1978, p. 124). A' in ignorance is intended to miss this pretense, to take S as speaking sincerely. But A, as part of the “inner circle” (to use Fowlers phrase), is intended to see everything the pretense, S's injudiciousness, A's ignorance, and hence Ss attitude toward S', A and what S' said. S' and A' may be recognizable individuals () or people of recognizable types”.

¹¹ “(...) satire and the forms of irony share many of the same features, especially the use of pretense and the necessity of multiple mental representations. (...) Although parody also requires the construction of multiple mental representations, it differs from satire because it relies on echoic mention and not pretense.”

texto fonte estabilizado; já a sátira se utiliza de partes originais de uma obra e procura verossímil com uma realidade externa.

2.2.2 Simpson (2003)

A partir de uma perspectiva discursiva, Simpson (2003) fornece uma definição linguística da sátira com um método, descrito nesta seção, para identificar e analisar suas diferentes instâncias. Teoricamente, a sátira é definida como uma prática discursiva que estabelece e resolve uma incongruência irônica entre um autor satírico com o propósito de criticar ou zombar um alvo satírico, podendo resultar em uma resposta humorística do público (JOHNSON; RIO; KEMMITT, 2010). Além disso, para ele, a sátira é superior ao que se conhece como gênero literário. Simpson (2003) afirma ainda que a sátira pode ecoar e fundir múltiplos discursos distintos do texto, pois os textos satíricos não dependem apenas da cultura, mas também são contextualmente dependentes. O meme "*Barbie Fascista*", ilustrada pela Figura 5, por exemplo, apareceu logo no final das eleições de 2018 como uma sátira antipetista e ganhou as redes sociais em poucos dias. Na imagem é possível observar uma boneca Barbie – branca e loira – e o enunciado "O PT destruiu a minha vida", insinuando que quem o profere é a Barbie. Ainda, por meio da ironia, é possível compreender que não é a boneca quem diz isso, mas sim a representação da mulher bolsonarista, a quem o meme faz uma crítica satírica. Nota-se, deste modo, que o reconhecimento do discurso satírico da *Barbie Fascista* é atingido a partir do contexto em que ele está inserido.

Figura 5 – Exemplo de sátira contextualmente dependente

"O PT quase destruiu a minha vida"



Fonte: Redes sociais do UOL.¹²

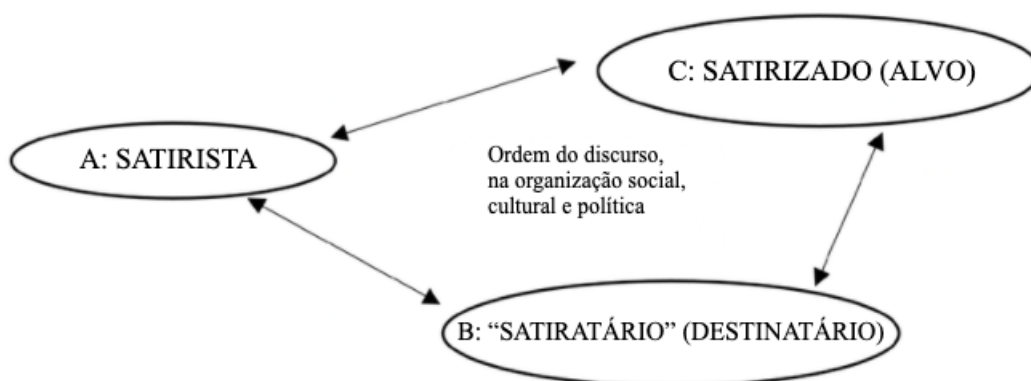
Segundo o autor, a sátira ocupa um lugar mais alto do que qualquer classificação linguística anterior dos conceitos: registro e gênero; e mais alto do que os críticos literários

¹² Disponível em: <<https://twitter.com/uol/status/1059767655463297025>>. Acesso em: 17 de out. 2022.

acreditam ser “gêneros literários”. Embora reconheça o papel das definições literárias, Simpson (2003) argumenta que a sátira é uma forma de humor que usa meios irônicos para atingir seus objetivos. Dessa forma, a sátira requer dois elementos principais: um gênero, definido como uma derivação em uma cultura particular, em um sistema de instituições e nas estruturas de crenças e conhecimento que envolvem e abrangem essas instituições e um registro que emana de uma desaprovação percebida, pelo satírico, de algum aspecto de um alvo satírico potencial (SIMPSON, 2003).

Como mostra a Figura 6, em seu modelo, Simpson apresenta uma tríade para a configuração da sátira, incluindo o que ele chama de “três posições do sujeito discursivo”, incorporando o *satirista* (o produtor), o *'satiratário'* (o destinatário)¹³ e o *satirizado* (o alvo) (SIMPSON, 2003, p.86). O autor aponta que o satirista e o satiratário são dois participantes que são “autênticos”, enquanto o satirizado não é bem-vindo no discurso satírico. Esse alvo pode ser um indivíduo, um evento, uma experiência ou até mesmo outro discurso. É importante mencionar que Simpson (2003) destaca que as posições de sujeito na tríade estão sujeitas a constantes deslocamentos e (re)organizações.

Figura 6 – Estrutura triádica da sátira como prática discursiva



Fonte: Adaptado de Simpson (2003, p.86) para o português.

Além disso, Simpson postula que uma sátira de sucesso mantém o princípio geral de entrega e recepção de humor ou pode encurtar a distância entre as posições A e B, vinculando assim a relação entre essas duas posições discursivas e alongar sua conexão com a terceira posição do sujeito C. No entanto, a sátira sem sucesso ou malsucedida pode alongar a conexão entre A e B, enquanto encurta a conexão entre os sujeitos B e C. Assim, a compreensão bem-sucedida da mensagem satírica resulta em uma interpretação

¹³ Como o esse sujeito destinatário, i.e., é o leitor/ouvinte que recebe o enunciado, utiliza-se a tradução do italiano (LAGHI, 2015) da adaptação de Simpson (2003).

humorística da mensagem por parte do público satírico, se o público simpatizar com o enunciado da mensagem.

Para o autor, o reconhecimento e a compreensão da sátira também incluem três estágios: o *primitivo*, a *dialética* e a *compreensão*. O primitivo é uma referência ecoica a um tipo de discurso, registro ou evento do gênero que pode ou não ser totalmente real. A dialética é onde um “elemento interno do texto” (SIMPSON, 2003, p. 89) colide de alguma forma com o primitivo e, assim, cria uma incongruência entre as expectativas alavancadas do conhecimento do ouvinte do primitivo (as notícias satíricas, por exemplo). Por último, a compreensão é onde ocorre a resolução da incongruência dialética (ou seja, o ouvinte interpreta a mensagem mais sutil e satírica). Em outras palavras, um texto satírico opera evocando um evento ou entidade de discurso anterior (o estágio primitivo) e, em seguida, produz um conflito de ideias internas ao texto que sinaliza uma incongruência (o estágio dialético) entre a forma do texto e a mensagem do texto. Desse modo, para demonstrar o que foi dito, retoma-se o exemplo (4), em que o “Mais Médicos”¹⁴ é o discurso anterior do discurso satírica, colidindo com o “Menos Médicos”, a incongruência presente na notícia satírica.

O conceito de incongruência é uma explicação predominantemente de como o humor é reconhecido e compreendido e tem sido usado como um veículo para analisar a estrutura linguística da sátira e para explicar como a compreensão do humor se desenvolve como um processo cognitivo (YOUNG et al., 2019; DYNEL, 2009; FORABOSCO, 2008; RITCHIE, 2005). Assim, o reconhecimento dessa incongruência, que requer conhecimentos específicos (cultural e contextual), é necessário para o terceiro estágio do processo satírico, a compreensão. Uma compreensão que resolve a incongruência entre o estágio primitivo e o estágio dialético resulta em humor.

A interpretação entre o primitivo e a dialética depende do acesso do satiratório a uma variedade de recursos de conhecimento, por exemplo, conhecimento geral e/ou universal e conhecimento de um evento ou área específica, o que resultam em uma captação satírica. Essa compreensão, de acordo com Simpson (2003), depende fortemente das reivindicações de validade universal de (HABERMAS, 1979)

O primitivo instanciado em um texto especificamente satírico funciona ao ecoar algum tipo de ‘outro’ evento de discurso, seja este outro gênero, dialeto ou registro, ou mesmo outra prática discursiva. Em contraste, a dialética é um elemento interno ao texto (em oposição ao intertextual), sendo normalmente posicionado após o primitivo, embora sua aparência pode ser isócrona (SIMPSON, 2003, p. 89 – tradução nossa)¹⁵.

No que diz respeito às propriedades linguísticas, a sátira funciona como um discurso

¹⁴ Disponível em: <<http://maismedicos.gov.br/>>. Acesso em: 15 de out. 2022.

¹⁵ “The prime instantiated in a specifically satirical text functions by echoing some sort of ‘other’ discourse event, whether that be another text, genre, dialect or register, or even another discursive practice. By contrast, the dialectic is a text-internal (as opposed to intertextual) element, which is normally positioned after the prime, although its appearance may sometimes be isochronous.”

de escalão superior, pois utiliza um gênero primário com o qual estabelece uma relação dialética¹⁶. Dessa forma, ativa um acontecimento anterior (real ou possível) que se torna um discurso ecoico do texto satírico – é a dissonância entre o domínio original e o dialético que cria uma estrutura pragmática para a interpretação. Logo, para reconhecer a manchete ilustrada no exemplo (5) como um discurso satírico, é necessário reconhecer o discurso ecoico presente no discurso anterior. Assim, o leitor precisa identificar o evento no qual o ex-presidente Temer, a partir de manobras políticas, assumiu a Presidência da República após o impeachment da ex-presidenta Dilma Rousseff¹⁷.

(5) Dilma entra com pedido no STF para ficar inelegível para cargos que tenham vice.

Assim como Kreuz e Roberts (1993), Simpson (2003) considera a ironia como um componente crucial para a criação da sátira e que o reconhecimento de uma peça satírica depende em última instância da realização de sua configuração irônica. Para o autor, é importante assimilar um dispositivo pragmático aos conceitos de primitivo e dialética, pois “esse dispositivo é a ironia, um ingrediente essencial do discurso satírico” (SIMPSON, 2003, p.90). Desse modo, o primitivo traz uma ironia ecoica, construindo um domínio de discurso estabelecido por um intertexto mediado, enquanto a dialética trabalha com a ironia em uma forma de oposição, manipulando o texto e criando contra expectativas no discurso.

Ser irônico pode servir a objetivos e funções comunicativas específicas (GIBBS, 2000), às vezes de maneiras mais eficientes do que a fala não irônica (KREUZ, 2000). A ironia é mais bem definida como uma diferença entre o que é dito e o que se quer dizer, com essa diferença levando o ouvinte a resolver aparentes incongruências em um enunciado. Attardo (2000) defende a ironia como inadequação relevante: o significado por trás de um enunciado irônico é indireto, mas o enunciado em si ainda tem alguma relevância para o contexto em que foi dito (GIBBS; COLSTON, 2007).

Logo, considerando o que já foi discutido sobre a sátira, paródia e ironia, a próxima seção descreve as notícias satíricas, que são um tipo de notícia apresentada que faz paródia ao noticiário verdadeiro, porém com conteúdo satírico – recorrendo ao humor, exagero e ironia – sobre eventos e acontecimentos atuais.

2.3 A SÁTIRA E AS NOTÍCIAS SATÍRICAS

As notícias satíricas são notícias fictícias que parodiam o gênero notícia e cobrem uma ampla escala de tópicos, incluindo questões sociais, política, entretenimento, esportes, entre outros. A maioria das notícias satíricas são baseadas em notícias verdadeiras e, por se tratar de uma obra construída pela sátira, são narradas através do absurdo, da

¹⁶ É por isso que Simpson (2003, p. 215), diferentemente da sátira, não vê a paródia como um gênero discursivo propriamente dito.

¹⁷ Disponível em: <<http://glo.bo/3EYC2A4>>. Acesso em: 15 de out. 2022.

graça e da incongruência e buscam criticar ou julgar acontecimentos da sociedade. Além disso, as notícias satíricas são frequentemente compartilhadas online e impactam como os indivíduos veem sua sociedade (RUBIN et al., 2016) e atravessam as fronteiras da mídia, ostentando uma grande variedade de canais e tipos – desde revistas a programas de TV, sites e personagens fictícios na Web (ERMIDA, 2012).

De acordo com Ermida (2012), a sátira presente em noticiário satírico online mantém o formato de notícias populares, como o layout das manchetes, a estrutura e o estilo linguístico das notícias, mas com um conteúdo que parodia a realidade. Conforme a Figura 7, é possível observar como a notícia verdadeira é construída pela:

- Manchete: “*Após divulgar reabertura das escolas para julho, secretaria de Educação de SP recua e diz que retorno não tem data definida*”.
- Lide: “*Anúncio tinha sido feito por meio de informativo enviado à imprensa e confirmado ao G1 por telefone. Em novo comunicado, pasta afirma apenas que volta às aulas será gradual e regionalizada*”.
- Nome do jornal com a data de publicação “*Por G1 SP*” e “*04/06/2020 13h11*”.
- Corpo da notícia.

Figura 7 – Exemplo de notícia real

Após divulgar reabertura das escolas para julho, secretaria de Educação de SP recua e diz que retorno não tem data definida

Anúncio tinha sido feito por meio de informativo enviado à imprensa e confirmado ao G1 por telefone. Em novo comunicado, pasta afirma apenas que volta às aulas será gradual e regionalizada.

Por G1 SP — São Paulo
04/06/2020 13h11 · Atualizado há 2 anos

Após anunciar a reabertura das escolas a partir de julho, a secretaria estadual de Educação voltou atrás e afirmou que o calendário de retorno às aulas ainda não tem data definida e depende de aprovação do Centro de Contingência do coronavírus de São Paulo.

As aulas presenciais na rede municipal e estadual estão **suspensas desde o final de março** por conta da pandemia do novo coronavírus.

Inicialmente, em e-mail enviado à Imprensa às 12h59, o governo informava que as aulas presenciais seriam retomadas em julho, de forma gradativa e regionalizada, podendo, ainda, ser antecipadas no caso das creches e unidades de educação infantil.

Em um novo e-mail enviado às 13h32 e às 15h36, porém, a informação havia sido retirada e, a data, negada.

Por outro lado, como mostra a Figura 8, a notícia satírica segue um modelo semelhante:

- Manchete: “Com reabertura das escolas, alunos levarão dever e vírus para casa”.
- Lide: “Os filhos têm permissão para tirar resultado negativo no teste”.
- Nome do jornal com a data de publicação: “Por Sensacionalista” e “Publicado em 31 jul 2020, 06h00”.
- Corpo da notícia
- Imagem ilustrativa.

Figura 8 – Exemplo de notícia satírica



Fonte: Sensacionalista.¹⁹

No entanto, apesar da semelhança entre as notícias, a notícia satírica faz uma crítica à reabertura das escolas no Estado de São Paulo durante a pandemia no ano de 2020 ao afirmar que além de levar dever, as crianças também levarão o vírus [da Covid] para casa, tendo em vista que ele é disseminado em ambientes com aglomeração de pessoas. Assim, o aspecto satírico surge quando “o substrato factual é comicamente estendido a

¹⁸ Disponível em: <<http://glo.bo/3TQgr2b>>. Acesso em: 20 de jan. 2022.

¹⁹ Disponível em: <<https://bit.ly/3RFVfQJ>>. Acesso em: 20 de jan. 2022.

uma construção fictícia onde se torna incongruente ou mesmo absurdo, de uma forma que cruza o entretenimento com a crítica.” (ERMIDA, 2012, p. 187 – tradução nossa).

Ermida (2012) ainda traz um modelo de notícias satíricas que abrange os mecanismos linguísticos que funcionam nesse exemplo de notícia. Esse modelo abrange e integra vários dos elementos da sátira, como a paródia e o humor, observados nas seções anteriores. O modelo de Ermida (2012) estipula que os três componentes principais e seus subcomponentes correspondentes sejam co-presentes em um texto para que ele se qualifique como uma notícia satírica.

Quadro 1 – Modelo de notícias satíricas

I	II	III
COMPONENTE INTERTEXTUAL	COMPONENTE CRÍTICO	COMPONENTE CÔMICO
I a) Componente estrutural		III a) Componente lexical
I b) Componente estilístico		III b) Componente pragmático
		III c) Componente retórico

Fonte: Adaptado de Ermida (2012, p. 194) para o português.

O **componente intertextual** consiste em uma base intertextual em relação a outras notícias, constituindo-se de dois subcomponentes: o **componente estrutural**: o texto deve fazer paródia do esquema organizacional e de desenvolvimento geral de uma notícia e o **componente estilístico**: o texto deve fazer paródia ao estilo de linguagem (estilo formal), a construção sintática da notícia (3^a pessoa e sentenças SVO) e o vocabulário (palavras objetivas) de uma notícia verdadeira. Quanto ao **componente crítico**, a notícia satírica deve possuir caráter de julgamento, enquanto desaprova, censura ou diminui aspectos da sociedade. Por último, assim como Rubin et al. (2016), a proposta desta pesquisa se concentra no *componente cômico*, que se refere à organização linguística responsável pela construção do efeito de humor no texto, fundamentando-se em três subcomponentes: o **componente lexical**: o texto deve ser organizado lexicalmente de modo que as palavras empregadas evoquem, acionem ou ativem *scripts* que são opostos e se sobrepõem; o **componente pragmático**: o texto deve ser construído pragmaticamente, de modo que a cultura e as referências extralinguísticas sejam importantes na interpretação do leitor e o **componente retórico**: o texto deve ser construído a partir de artifícios retóricos (como a antítese ou hipérbole) para instanciar/intensificar as oposições de *scripts*.

¹⁹ “[...]the factual substratum is comically extended to a fictitious construction where it becomes incongruous or even absurd, in a way that intersects entertainment with criticism.”

Com relação às denominações do que é uma sátira noticiosa, a questão terminológica não é um consenso, pois trabalhos que abordam o mesmo tema – os textos humorísticos que fazem paródia às notícias verdadeiras – estabelecem a semelhança com o gênero notícia, porém utilizam nomenclaturas distintas, como: *desnotícias*, (ROCHA, 2017; FIGUEIRA, 2018), *notícias humorísticas* (SILVEIRA, 2019) e *pseudonotícias* (MORETT, 2015). Nesta pesquisa utiliza-se a denominação do termo notícias satíricas, assim como Souza (2013).

Souza (2013) ressalta o objetivo comunicativo de criticar, converter e moralizar das notícias satíricas. O autor aponta que estas três ações estão interligadas ao funcionamento da notícia satírica, uma vez que há “uma crítica bem-humorada em relação a algo que se encontra em desconformidade com a moral prezada na comunidade e a crítica é realizada para converter aquilo imoral em moral, ou seja, tem por objetivo moralizar” (SOUZA, 2013, p. 155). Além disso, considerando a intertextualidade desse tipo de notícia, Souza (2013) ainda afirma que as notícias satíricas também apresentam características de outros gêneros. Para ele, é possível perceber uma apropriação de estruturas do gênero notícia utilizadas para passar informações aos indivíduos da sociedade, mas de um modo subversivo, característico das notícias satíricas. Além disso, em referência ao caráter ficcional das notícias satíricas, Souza (2013) sinaliza que:

a montagem da imagem ilustrativa, a utilização de fontes fictícias e a elaboração de depoimentos fictícios de acordo com estas fontes, a mescla entre o estilo jornalístico de notícia e o estilo satírico de composição, bem como a própria desfiguração da realidade realizada por meio da inclusão de elementos fictícios para a construção da ‘realidade fictícia’ que é noticiada (SOUZA, 2013, p.159).

Esse tipo de fusão entre os elementos ficcionais citados que compõem a notícia satírica e a notícia real causa uma espécie de distanciamento do que normalmente o gênero notícia realiza. Dessa forma, a notícia satírica possibilita também uma nova interpretação (SOUZA, 2013).

Já Rocha (2017, p.74) considera as notícias satíricas como um “gênero midiático híbrido entre o gênero opinativo jornalístico e o entretenimento midiático”. Desse modo, essas notícias transitam por esses gêneros, atravessando tanto o universo da realidade, como o da ficção, e mantém uma relação entre o humor e o jornalismo de opinião. O autor se utiliza da nomenclatura *desnotícia*, pois acredita que ao inserir o prefixo ‘*des*’ na palavra notícia, torna-se evidente a sua desconstrução como notícia. Estendendo a proposta da notícia satírica como uma desconstrução, Figueira (2018) estabelece que esse tipo de notícia é definido como um gênero do discurso recente que mimetiza as notícias por meio de paródias e também as denominam como *desnotícias*. Apesar de as *desnotícias*, em um primeiro momento, possuir uma semelhança com o gênero textual notícia, a partir do momento em que o leitor aprofunda sua leitura, ele observa um estranhamento em relação às informações e aos fatos, uma vez que contêm elementos humorísticos. O autor salienta

que os estudos são recentes e por isso ainda não existe consenso sobre sua denominação (FIGUEIRA, 2018).

Dessa forma, os textos de notícias satíricas e humorísticas são paródias, ou seja, possuem elementos intertextuais que parafraseiam outro texto, introduzindo um sentido diferente do original. Nesse contexto, é importante que o leitor recorra a sua memória ao gênero notícia, a respeito do campo jornalístico e a assuntos relacionados ao cotidiano da sociedade em que está inserido, pois somente dessa maneira será possível perceber o efeito humorístico. Assim, entende-se também como ocorre a subversão dos fatos (FIGUEIRA, 2018).

No entanto, pensando pelo viés do humor, segundo Silveira (2019), as notícias humorísticas também são caracterizadas por desconstruírem fatos verificáveis com a intenção de criticar a sociedade e causar humor e não apenas informar. A autora ainda afirma que elas têm um local fixo para serem publicadas, fazendo com que alcance muitos leitores. Segundo Silveira (2019), as notícias humorísticas apresentam as seguintes características:

- a) pertencem tanto à prática social do jornalismo, pois a simulam, quanto à prática social do humor;
- b) têm como propósito comunicativo não só o entretenimento, mas também, e significativamente, a crítica social;
- c) têm um alcance significativo no mundo virtual, no que diz respeito à distribuição e ao consumo, verificados por meio de curtidas, compartilhamentos e comentários no site, sempre que possível, e na página do Facebook;
- d) são produzidas a partir da estrutura do gênero notícia;
- e) a narração e a argumentação caracterizam a produção dos textos das notícias humorísticas em relação à presença dos pré-gêneros;
- f) são um gênero situado, pois elas são uma forma bem específica de texto de humor, por serem notícias que representam desconstrução da realidade e de teor humorístico.
Além disso, são sócio-historicamente contextualizadas, por terem um local próprio de publicação e por se valerem da intertextualidade.
- g) a intertextualidade é o principal recurso para se produzir as notícias humorísticas, pois elas têm como base fatos veiculados pelos portais de jornalismo;
- h) são um gênero híbrido, já que há a apropriação da prática do jornalismo pela prática humorística (SILVEIRA, 2019, p.43)

Morett (2015), no que lhe concerne, aponta que a nomenclatura é uma questão delicada, dado que não existe ainda um nome designado para esse tipo de texto de humor. Assim como nesta pesquisa, o estudo da autora é pautado nos textos do Sensacionalista, porém denomina as notícias como ‘pseudonotícias’ e salienta que o gênero ‘pseudojornal’ é uma paródia de notícias e que sua base é a crítica e o humor (MORETT, 2015). Nas palavras da autora, as ‘pseudonotícias’ são “paródias das tradicionais notícias jornalísticas, imitando toda sua estrutura, composição e lógica, mas com conteúdo fantasioso e humorístico” (MORETT, 2015, p.19). Morett (2015) afirma ainda que as ‘pseudonotícias’ englobam

diferentes assuntos, fazendo referências muito distantes, requerendo um alto grau de contextualização do leitor.

Finalmente, ao observar as investigações do gênero humorístico aqui abordadas, é possível perceber que as notícias humorísticas têm como base fatos que já foram produzidos nos portais tradicionais do jornalismo. É significativo, portanto, que o leitor tenha um bom conhecimento contextual e intertextual para a compreensão do sentido, porque somente dessa maneira irá reconhecer as críticas e o teor humorístico que o texto deseja transmitir.

Entende-se, portanto, que nem sempre o leitor conseguirá atingir o efeito humorístico e fictício das notícias satíricas. Assim, na próxima seção será abordado porque esse tipo de notícia pode ser considerada um conteúdo enganoso.

2.4 AS NOTÍCIAS SATÍRICAS COMO CONTEÚDO ENGANOSO

Nas duas últimas décadas, a utilização de mídias sociais vem mudando a estrutura de reprodução de notícias *on-line* que são veiculadas na rede. Um conteúdo jornalístico pode ser transmitido de uma pessoa para a outra sem uma controle ou uma filtragem editorial (ALLCOTT; GENTZKOW, 2017). Esse novo ambiente informacional, que possibilita um maior alcance de diferentes tipos de usuários em uma velocidade rápida, facilita a reprodução de conteúdos equivocados ou falsos, comumente chamados de *Fake News*.

Apesar da expressão *Fake News* ter se popularizado nos últimos anos e estar nos holofotes devido ao seu impacto nas questões fundamentais da sociedade, seu conceito não é algo tão novo. Desde a Antiguidade, formas de escrita eram encontradas impressas em pedras, argila e papiro. Era por meio do controle da informação, verdadeira ou falsa, que os líderes de uma sociedade demonstravam – e ainda demonstram – o seu poder. Ao mesmo tempo, aqueles que não mantinham a autoridade sobre o povo, mas tinham o conhecimento da palavra, utilizavam-se da informação para criticar ou ridicularizar o governo e seus representantes.

Em 1274 a.C., o faraó Ramsés II do Egito, deixou registrada em papiros e monumentos de pedras o seu indiscutível triunfo sobre os hititas de Cades. A gloriosa conquista egípcia foi contestada quando cartas privadas entre Ramsés e o rei Hitita Hattusili III foram descobertas, nas quais tratavam a batalha em Cades como uma vitória. Quinze anos após a batalha, um tratado mostrou os dois lados se reconhecendo como iguais (ANG; ANWAR; JAYAKUMAR, 2021).

No século VI, de acordo com o historiador Darnton (2017), a reputação do imperador Justiniano foi arruinada pelo texto “História Secreta” (*Anedokta*, no título original), escrito pelo historiador bizantino Procópio de Cesareia (aproximadamente 500-554 a.C.). No século XVI, em 1522, Pietro Aretino, jornalista da época, escrevia poemas curtos e satíricos para criticar e desacreditar, principalmente, a igreja e seus cardeais. Tais poemas e sonetos eram fixados na estátua de Pasquino, em Roma, e acabaram ficando conhecidos

como ‘pasquinadas’ (DARNTON, 2017). As pasquinadas de Aquino deram origem a uma variedade de notícias falsas francesas no século XVII – os *canards*, podendo ser utilizadas para designar um boato ou uma história infundada (BURKHARDT, 2017).

Burkhardt (2017) explicita que no século XVIII, em 1710, Jonathan Swift reclamou de falsas notícias políticas em seu ensaio *The Art of Political Lying*. Ainda no mesmo século, em 1770, os chamados ‘homem-parágrafo’ recolhiam fofocas e as escreviam em um único parágrafo. Estas fofocas eram vendidas e impressas em forma de pequenas reportagens – na maioria das vezes difamatórias.

Foi no século XIX que os jornais com informações apuradas, como os que conhecemos atualmente, popularizaram-se. Conforme Abad (2019), em 1835, o jornal nova-iorquino *The Sun*, informou que um cientista britânico havia avistado, graças a seu potente telescópio, vida extraterrestre na Lua. A notícia se espalhou pelos Estados Unidos e Europa, chamando a atenção dos leitores da época. Quando descoberta, a farsa ficou conhecida como ‘A Grande Mentira da Lua’.

Após a criação do rádio, no século XX, a disseminação das notícias se tornou mais rápida e mais passível de manipulação ou engano. Em 1926, na Inglaterra, o noticiário *Broad Casting the Barricade*, do padre Ronald Knox causou uma grande confusão ao fazer uma paródia e informar que Londres estava sendo atacada por comunistas, o Parlamento estava sitiado e o Big Ben havia sido explodido. Em um contexto muito semelhante, em 1938, uma transmissão de rádio causou pânico aos estadunidenses ao narrar a obra ‘A Guerra dos Mundos’ (publicada em 1898). Ouvintes que haviam sintonizado tardiamente acreditaram ser verídico o relato da invasão dos marcianos (BURKHARDT, 2017).

Finalmente, no final do século XXI, a utilização das mídias sociais permitiu a mudança da estrutura de reprodução de notícias on-line compartilhadas na rede. A explosão de informações veiculadas nesses meios desencadeou o processo de divulgação do conteúdo jornalístico, transmitido de uma pessoa para a outra, sem um controle ou uma filtragem editorial. Esse novo modo de disseminação da informação possibilitou um maior alcance de diferentes tipos de usuários em uma velocidade cada vez mais rápida. Assim, o ambiente virtual se tornou ideal para a divulgação vertiginosa de notícias falsas.

Assim, a capacidade de divulgação das notícias falsas está intrinsecamente ligada aos suportes de cada época, como os papiros e pergaminhos na Antiguidade, a criação da imprensa no século XV e o rádio e a TV nos séculos XIX e XX, respectivamente. No século XXI, com a Internet, a velocidade da comunicação cresceu e acelerou o tempo de disseminação das notícias, atingindo o maior número de pessoas em todo lugar do planeta em um tempo muito pequeno.

As redes sociais como novos meios de comunicação permitiram um maior espaço para o indivíduo expor seus pensamentos, opiniões, emoções e posições de suas ideias, o que fortaleceu um comportamento individualista. Assim, o ambiente digital se tornou, com o passar dos anos, o habitat ideal para a disseminação da desinformação, pois é a

partir das atuais tecnologias que uma notícia falsa é desenvolvida e compartilhada de uma forma sistêmica. O excesso de informação veiculada nas redes influenciou a sociedade atual, impactando diretamente em sua forma de pensar e agir no mundo contemporâneo.

Para Boarini e Ferrari (2021, p. 40), “uma vez que o processo de desinformação se torna endêmico e robusto, fica garantida a eficácia tanto na entrega da informação como na obtenção da instabilidade com ela pretendida”. Um dos principais fatores para o processo de desinformação é como a configuração das mídias sociais se comporta para uma reconfiguração da sociedade como um todo, tornando-a imediatista, com enorme alcance, mas que, ao mesmo tempo, insiste em dialogar entre algumas bolhas (BURKHARDT, 2017).

No estudo realizado por (VOSOUGHI; ROY; ARAL, 2018), as *Fake News* se espalham 70% mais rapidamente que as notícias verdadeiras, mostrando ainda que as notícias falsas são mais inusitadas do que as reais, o que indica que os usuários estão mais dispostos a compartilhar esse tipo de informação. Para a pesquisa do *Pew Research Center*²⁰, 64% dos americanos acreditam que as informações falsas se confundem com fatos básicos das questões do cotidiano. Além disso, 39% se sentem confiantes para identificar uma notícia falsa, 23% dos americanos já compartilharam notícias falsas e 14% afirmam saber que o conteúdo não era verdadeiro.

No ano de 2017, conforme a publicação do jornal *The Guardian*²¹, após pesquisadores monitorarem mais de 4,5 bilhões de palavras na Internet e nas redes sociais e constatarem a propagação de mais de 365% do termo, principalmente após as eleições norte-americanas em 2016, o termo ‘fake news’ foi considerado a palavra do ano pela editora Collins (FLOOD, 2017). A Figura 9 apresenta o interesse ao longo do tempo²² em busca no Google da expressão ‘fake news’ em todo o mundo no período entre 1 de janeiro de 2011 a 31 de dezembro de 2021. Em janeiro de 2011 a popularidade do termo atingiu a pontuação 3. Em outubro de 2016, período próximo às eleições dos EUA (e quando o termo começou a se popularizar) a popularidade atingiu 6 pontos. Em março de 2020, quando a OMS decreta a pandemia de Covid, a busca pelo termo chega em seu ápice, atingindo 100 pontos. E, em 31 de dezembro, o termo atingiu 21 pontos de popularidade.

²⁰ Disponível em: <<https://pewrsr.ch/3DjWwTx>>.

²¹ Disponível em: <<https://bit.ly/3QkvRZy>>.

²² De acordo com o Google, os números que representam o interesse de pesquisa relativo ao ponto mais alto no eixo y. Assim, um valor de 100 representa o pico de popularidade de um termo, um valor de 50 significa que o termo teve metade da popularidade e uma pontuação de 0 significa que não havia dados suficientes sobre o termo.

Figura 9 – Busca do termo fake news no Brasil entre 01/01/2011 a 31/12/2021



Fonte: Elaborado pela autora.

A partir de então, o termo entrou para o vocabulário do mundo todo, não só pelo conhecimento do aumento da divulgação desse tipo de notícia, mas também pela compreensão dos danos que uma notícia falsa poderia causar à comunidade. Entretanto, muitos grupos de políticos, líderes e pessoas mal-intencionadas criaram e espalharam informações mentirosas para desacreditar a imprensa tradicional. As ‘*Fakes News*’ estiveram presentes nas eleições norte-americanas de 2016, na disputa da presidência do Brasil em 2018, nos debates do Brexit, na pandemia da Covid-19 e mais recentemente nas eleições brasileiras em 2022.

Seria possível entender o conceito de ‘*Fake News*’ em apenas uma definição? A heterogeneidade do termo e suas diversas definições podem ser um grande problema para os estudiosos e curiosos do tema, pois é um termo que passou a significar itens diferentes para pessoas diferentes. Entretanto, o que se entende transpassa as limitações conceituais, pois uma notícia pode ser projetada intencionalmente para enganar o leitor, ser criada para atrair cliques e obter lucro ou ser notícias satíricas com o objetivo de entreter o público. Grosso modo, o termo ‘*Fake News*’ pode ser definido como notícias imprecisas e, muitas vezes, fabricadas intencionalmente (QUANDT et al., 2019).

É comum, especialmente nas mídias sociais, rotular um conteúdo como falso quando certos fatores exigem vigilância. Esses conteúdos podem ser, por exemplo, publicações criadas nas redes sociais, imagens contendo texto duvidoso ou links que direcionam para possíveis notícias ou outros elementos utilizados para enganar e não informar. Dessa forma, nota-se que ao fazer uma tradução literal do termo usado para identificar esses casos, o termo ‘*Fake News*’ foi traduzido para notícias falsas, mas todas as notícias dentro desse contexto podem ser encaixadas nessa terminologia? Assim como os cientistas do grupo *Social Media, Online Disinformation and Elections*²³, esta pesquisa acredita que o termo ‘*Fake News*’ é inadequado, pois se trata de diversos fenômenos, como notícias satíricas, e não apenas um conteúdo falso fabricado intencionalmente.

De acordo com Rubin, Chen e Conroy (2015), existem três tipos tradicionais de

²³ Disponível em: <<https://gate-socmedia.group.shef.ac.uk>>

conteúdo enganoso (do inglês, *deception*):

- i) **Notícias fabricadas** – normalmente produzidas pelo que é chamado de imprensa marrom ou tabloides;
- ii) **Boatos** – são consideradas aquelas notícias disfarçadas para enganar o público e podem ser divulgadas por descuido pelas agências de notícias tradicionais e
- iii) **Notícias satíricas** – são notícias parecidas com as notícias reais, porém, são criadas para fins de humor. Se não reconhecidas como tal, notícias satíricas podem criar dificuldades de entendimento e falsas crenças nas mentes dos leitores.

Pensando na descrição e no processamento automático desse tipo de conteúdo enganoso, os autores ainda sugerem que um *corpus* de notícias falsas deve atender aos seguintes requisitos:

1. **Alinhamento de notícias falsas com verdadeiras:** os métodos preditivos devem conseguir encontrar padrões e regularidades em dados positivos e negativos. Sobre este requisito, o *subcorpus* utilizado nesta tese (cf. Capítulo 4), está alinhado em notícias satíricas e notícias reais.
2. **Notícia em formato textual:** o formato textual ainda é o meio preferido para análise em PLN. Caso a notícia venha acompanhada de imagem ou vídeo – ou que estejam nesses formatos (imagem e/ou vídeo) – ela deve ser transcrita para o formato textual. Em relação a este, o *subcorpus* se encontra em formato textual.
3. **Verificação da verdade:** deve ser necessária possibilidade da veracidade das notícias, ou seja, se ela realmente é genuína, ou não. A respeito deste requisito, procurou-se coletar as notícias reais apenas de veículos tradicionais de comunicação.
4. **Homogeneidade do tamanho dos textos:** os textos devem ter tamanhos semelhantes. Por exemplo, um resumo de um parágrafo do Facebook e um editorial longo não constituem um conjunto de dados homogêneo. Desse modo, pode ser realizada a normalização para alguns conjuntos de dados desiguais. Quando a homogeneidade do tamanho dos textos, para a construção do *subcorpus* procurou-se não seguir este requisito, visto que a diferença de tamanho pode indicar se é ou não uma notícia satírica.
5. **Homogeneidade na forma escrita:** os textos devem estar alinhados em relação aos tipos de notícias (por exemplo, notícias de última hora, editoriais, artigos de opinião) e tópicos (por exemplo, negócios, política, ciência, saúde), usar tipos parecidos de autores (por exemplo, treinados profissionalmente, pessoas comuns *versus* jornalistas ou sério *versus* bem-humorado) e também podem ser comparados entre os meios de

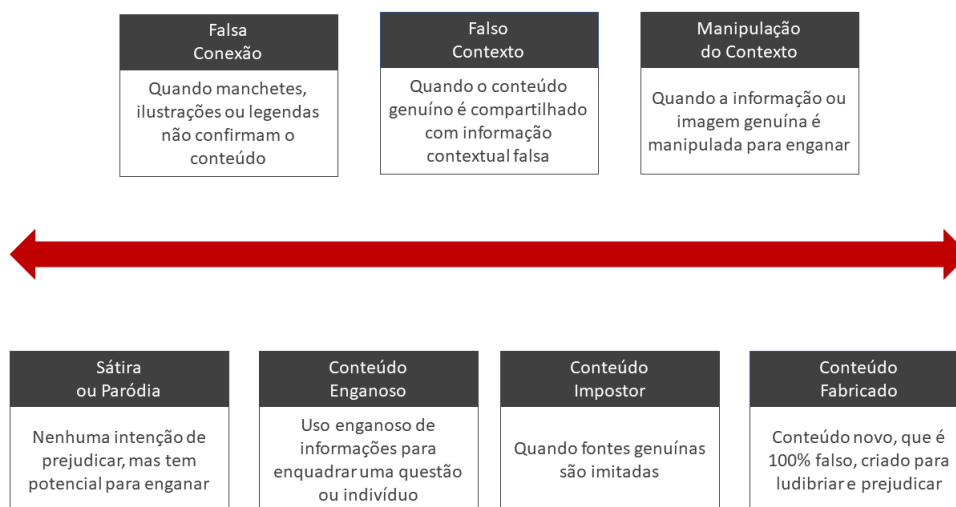
comunicação. Sobre a homogeneidade na forma escrita, tanto as notícias satíricas quanto as reais são do mesmo tópico (política).

6. **Janela de tempo predefinida:** o *corpus* deve ser coletado em um prazo de tempo já definido, uma vez que pode haver uma variação na escrita de um determinado tópico nos últimos 2–3 anos. Em relação à janela de tempo predefinida, estabeleceu-se um período de dois anos (entre 2016 e 2018).
7. **Modo que as notícias são entregues:** o conhecimento de como a notícia é entregue cria contexto para a interpretação em uma determinada situação (por exemplo, humor; noticiabilidade, credibilidade; absurdo; sensacionalismo). No que diz respeito ao modo que as notícias são entregues, considera-se apenas as notícias satíricas para análise do conteúdo enganoso.
8. **Preocupações pragmáticas:** incluem custos de direitos autorais (*copyright*), disponibilidade pública, facilidade de obtenção, volume geral adequado de dados, graus de divulgação e privacidade dos escritores, entre outros fatores. Aqui, espera-se que tanto o *corpus* quanto o *subcorpus* possa ser disponibilizado.
9. **Língua e cultura:** é necessário definir qual o idioma das notícias, pois, técnicas de detecção de notícias falsas para uma determinada língua podem não ser válidas para outras. Por fim, em relação à língua e cultura, o *subcorpus* foi construído para português do Brasil.

Além disso, independentemente da aplicação, sua descrição linguística, identificando-se os mecanismos de construção de uma sátira, mostra-se como uma tarefa importante para a Linguística, sobretudo ao PLN e na detecção automática de conteúdo enganoso.

Wardle e Derakhshan (2017) afirmam que para entender o ecossistema das informações que veiculam nas mídias sociais atualmente, descrito na Figura 10, é necessário compreender que os diferentes tipos de conteúdo que estão sendo criados e compartilhados, as motivações de quem cria esses conteúdos e as formas como esse conteúdo está sendo disseminado.

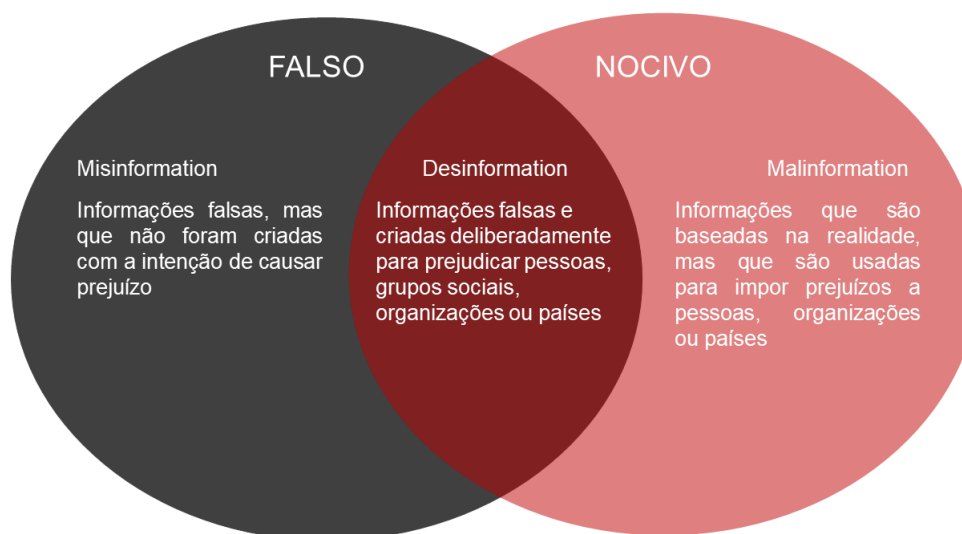
Figura 10 – Ecosystema da desinformação



Fonte: Adaptado de Wardle e Derakhshan (2017) para o português.

Os autores ainda definem esses conteúdos enganosos em três conceitos: *misinformation*, *disinformation* e *malinformation*. Ao passo que a *misinformation* é a informação falsa compartilhada de forma imprudente, mas sem a intenção de causar prejuízo, a *disinformation* é a informação falsa, criada propositalmente para causar danos às pessoas, ao país ou até mesmo a grupos sociais e organizações. Por último, a *malinformation* é a informação baseada na realidade, mas que é usada para impor prejuízos, como mostra a Figura 11. Esta distinção é importante para delimitar que nem todo conteúdo enganoso tem a intenção de enganar o leitor, como é o caso das notícias satíricas.

Figura 11 – Desordem da desinformação



Fonte: Adaptação de Wardle e Derakhshan (2017) para o português.

Tandoc, Lim e Ling (2018) apontam a existência de seis tipos de conteúdos enganosos:

- i) **Sátira** – são as notícias que focam no humor e no exagero como forma de informar o leitor sobre eventos verdadeiros que incluem política, economia ou temas sociais;
- ii) **Paródia** – são notícias que também buscam o humor. Ao contrário das notícias satíricas, essa categoria lida com conteúdo não factual, ou seja, o conteúdo da paródia em forma de notícia é geralmente fabricado;
- iii) **Notícias fabricadas** – são as notícias apresentadas no estilo narrativo dos veículos tradicionais de imprensa e que se baseiam em assunto sem nenhuma base factual;
- iv) **Manipulação de imagens ou vídeos** – com o crescente acesso a ferramentas de compartilhamento e edição, muitas notícias falsas utilizam-se da manipulação de fotos e vídeos para dar veracidade a suas narrativas;
- v) **Materiais publicitários disfarçados de notícias genuínas** – são notícias que, aparentemente, têm o mesmo formato de um conteúdo produzido pelas grandes mídias jornalísticas e de comunicação, mas, na verdade, são feitas por empresas de publicidade ou de relações públicas que buscam promover clientes e/ou produtos e
- vi) **Propaganda** – são as notícias criadas por uma entidade política com o objetivo de influenciar a opinião da população sobre o governo, organizações e lideranças.

Segundo os autores, o conteúdo enganoso ainda podem ser classificadas em outras duas dimensões: a facticidade do seu conteúdo e a intenção do produtor da notícia (TANDOC; LIM; LING, 2018). A facticidade refere-se ao grau de como as notícias falsas lidam com fatos. Desse modo, uma sátira pode retratar um conteúdo factual através da crítica, da sátira e do humor, alterando o formato de sua apresentação. Entretanto, as notícias como paródia ou as notícias fabricadas tratam de diferentes temas sociais, porém remodelam a sua facticidade. Já as notícias fabricadas não estão amparadas em acontecimentos reais.

A segunda dimensão corresponde à intenção do autor, relacionando-se ao grau com que os autores procuram enganar o público. Os autores das notícias satíricas e de paródias, por exemplo, lidam com um objetivo parecido, uma vez que a sua intenção é provocar um sentido de humor. Contudo, os autores das notícias fabricadas têm como propósito enganar deliberadamente o leitor.

Quadro 2 – Facticidade e intenção do texto

	Intenção do autor	
Nível de facticidade	Alto	Baixo
Alto	Publicidade e propaganda	Notícias satíricas
Baixo	Notícia fabricada	Notícia paródica

Fonte: Adaptado de Tandoc, Lim e Ling (2018, p. 12) para o português.

O Quadro 2 descreve a formatação textual proposta por (TANDOC; LIM; LING, 2018) de acordo com a sua intenção. Conforme este modelo descrito, os autores procuram avaliar por meio de duas variáveis: o nível de facticidade e a intenção do autor em enganar. Desse modo, as notícias satíricas, por exemplo, possuem um alto índice de facticidade, mas baixa intenção de enganar, à medida que a notícia paródica possui um baixo nível de facticidade e uma baixa intenção de enganar.



Neste capítulo foram apresentados os principais conceitos que envolvem a sátira, discorrendo sobre sua origem e sua percepção não apenas como um gênero literário, mas sim como um modo literário multifacetado que pode se manifestar em diversos formatos e gêneros. A partir da conceitualização da sátira, também foi discutido como a sátira pode ser estudada pelo olhar linguístico, vista como prática discursiva – e não como um gênero de discurso. Também foi destacado como ela se articula com o humor, a ironia e com a paródia, elementos essenciais para a análise no Capítulo 5.

Finalmente, foram introduzidas também denominações que norteiam as notícias satíricas e como e o porquê delas se constituírem como conteúdo enganoso. É importante ressaltar que a intenção do conteúdo satírico não é enganar o leitor e, desse modo, uma compreensão equivocada da mensagem satírica é um efeito indesejado da sátira. Assim, a compreensão das notícias satíricas como uma notícia falsa, mesmo sem ter a intenção de enganar o leitor, é um dos pilares que norteiam esta pesquisa, uma vez que a criação de recursos como o proposto nesta tese pode auxiliar na detecção automática conteúdo enganoso.

3 TRABALHOS RELACIONADOS

Neste capítulo é apresentada uma revisão geral de trabalhos relacionados à detecção de notícias satíricas em textos. O capítulo é organizado em duas categorias, conforme cada proposta ou abordagem que cada trabalho utiliza: i) os principais métodos computacionais, em que são mostrados os trabalhos que usam abordagens computacionais da análise de notícias satíricas e ii) as características linguísticas de notícias satíricas, onde são descritos os principais padrões linguísticos extraídos a partir de estudos de PLN.

3.1 MÉTODOS DE PLN VOLTADOS ÀS NOTÍCIAS SATÍRICAS

No capítulo anterior, discutiram-se as divisões das notícias falsas segundo Rubin, Chen e Conroy (2015), Wardle e Derakhshan (2017) e Tandoc, Lim e Ling (2018) e em todas as abordagens as notícias satíricas são consideradas um tipo de conteúdo falso, visto que elas podem ser prejudiciais por serem potencialmente enganosas quando não identificados os elementos figurados presentes no texto. Assim, esse contexto gerou o interesse não só de linguistas, mas também de cientistas da computação para sua detecção automática ou então a possibilidade de diferenciá-las de notícias verdadeiras. Como o foco desta tese é propor um mapeamento de recursos linguísticos presentes em notícias satíricas, nesta seção é apresentada uma revisão dos principais trabalhos sobre a detecção automática de conteúdo satírico que antecederam esta pesquisa. Ressalta-se que embora a maioria das pesquisas aqui retratadas seja para o processamento do inglês (BURFOOT; BALDWIN, 2009; RUBIN et al., 2016), há trabalhos para o alemão (MCHARDY; ADEL; KLINGER, 2019), espanhol (SALAS-ZÁRATE et al., 2017; BARBIERI; RONZANO; SAGGION, 2015), francês (IONESCU; CHIFU, 2021; LIU et al., 2019), romeno (ROGOZ; MIHAELA; IONESCU, 2021), turco (TOÇOĞLU; ONAN, 2019), árabe (SAADANY; ORASAN; MOHAMED, 2020), português europeu (CARVALHO et al., 2020), português do Brasil (WICK-PEDRO et al., 2020) e multilíngue (ABONIZIO et al., 2020; GUIBON et al., 2019).

É importante ressaltar que apesar do uso de aprendizado de máquina (AM) ou das redes neurais (RN) não é esse o foco ou objetivo deste trabalho, espera-se que esta tese possa ser um subsídio de recursos linguísticos para detecção automática de notícias para o português do Brasil. Assim, mesmo não utilizando estes recursos computacionais, é prudente abordar trabalhos sobre métodos de AM ou RN para detecção automática de sátira e humor em textos jornalísticos ou, em alguns casos, redes sociais.

Nesse contexto, alguns trabalhos mostram ser possível apreender habilidades linguísticas do humor presentes no enunciado satírico. Um exemplo é o estudo de Sarkar, Yang e Mukherjee (2018), que investigaram uma abordagem de rede neural profunda robusta e hierárquica para detecção de sátira, capaz de capturar sátira tanto no nível da sentença quanto no nível do documento. Os autores mostraram que o modelo apresentado pode

capturar a sátira de forma mais eficiente do que modelos já existentes, usando apenas *word embeddings*¹ pré-treinados como entrada, sem o auxílio de qualquer informação sintática ou recursos linguísticos. Em seguida, foi realizada uma extensa comparação com vários métodos de RN de última geração para detecção de notícias satíricas que também foram exploradas em um conjunto de dados de notícias de sátira de assuntos reais. Por fim, uma análise dos modelos aprendidos revelou a existência de algumas sentenças-chave, como a última sentença, são importantes para detectar a sátira. Esse resultado se relaciona com o trabalho de Rubin et al. (2016), que afirmam que a última sentença evidencia um absurdo ou introduz um novo elemento na história, provocando o efeito de humor.

Em outra perspectiva, o trabalho de Zhang et al. (2020), inspirado na expressão “*birds of a feather flock together*”², mostra um novo método, que não utiliza nem rede neural e nem extração de características linguísticas para a classificação de notícias satíricas. Segundo os autores, o método proposto é computacionalmente eficaz, porque os modelos de linguagem entre documentos de notícias satíricas e notícias verdadeiras são sensíveis quando aplicados a documentos fora de seus domínios. Os autores acreditam que como as notícias satíricas geralmente são compostas por histórias com conteúdo absurdo, é fácil para as pessoas com conhecimento e formação cultural correspondentes reconhecê-las. Desse modo, acreditam que uma representação semelhante a de como as pessoas discernem informações “ilógicas” presentes nas notícias satíricas pode alavancar modelos computacionais que têm a capacidade de obter conhecimento de domínio para avaliar um texto satírico como seres humanos. Assim, ao alavancar as pontuações surpresa de diferentes modelos de linguagem, as notícias satíricas foram diferenciadas de notícias verdadeiras de forma eficaz. Este método não é apenas livre de extração de numerosas características linguísticas, como os trabalhos que abordam a detecção de notícias satíricas, mas também não requer estruturas de rede neural sofisticadas ou *word embeddings* avançados. O método de Zhang et al. (2020) supera métodos anteriores (RUBIN et al., 2016; YANG; MUKHERJEE; DRAGUT, 2017; SARKAR; YANG; MUKHERJEE, 2018), alcançando uma acurácia³ de 97,97% e uma precisão⁴ de 94,55% no conjunto de dados de validação e 96,82% de acurácia e 93,67% de precisão no conjunto de dados de teste.

Shabani e Sokhn (2018) abordaram a detecção de notícias falsas e satíricas propondo um método de aprendizado de máquina híbrido. Segundo os autores, esse sistema combina o fator humano com a abordagem de aprendizado de máquina e um modelo de tomada de decisão que estima a confiança de classificação dos algoritmos e decide se a tarefa precisa de

¹ As *word embeddings* são vetores densos que representam palavras dentro de um espaço latente. Essas incorporações são geralmente aprendidas a partir de tarefas genéricas não supervisionadas, como a previsão da próxima palavra.

² Em português, seria equivalente a “cada qual com seu igual”.

³ O termo acurácia indica uma performance geral do modelo, como quantas classificações o modelo classificou corretamente.

⁴ A precisão é uma métrica definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos.

entrada humana ou não. O objetivo era distinguir sátira ou paródia e conteúdo fabricado usando o conjunto de dados público de conteúdo falso *versus* satírico. Para essa tarefa, foram aplicados modelos de aprendizado de máquina para classificar automaticamente as notícias como falsas ou satíricas e, em seguida, foram identificados recursos que podem melhorar a precisão. No entanto, devido à dificuldade da classificação demandar de checagem de fatos, o *crowdsourcing* foi utilizado como um serviço para obter melhor precisão, uma vez que os humanos deveriam classificar os artigos de política e relacioná-los como histórias falsas ou satíricas. O interessante do trabalho de Shabani e Sokhn (2018) é que a abordagem proposta fornece maior precisão a um custo e latência aceitável, pois combina a eficácia dos algoritmos de aprendizado de máquina com o conhecimento humano, por meio da aplicação de *crowdsourcing* nos casos em que os algoritmos de aprendizado de máquina não conseguem executar com alta precisão. Conforme concluem os autores, a abordagem híbrida aumenta a precisão geral em até 87%.

A pesquisa de Horvitz, Do e Littman (2020) se destaca por compreender que as notícias satíricas exigem uma apreciação de um contexto verdadeiro e não humorístico ao propor uma abordagem em que primeiro se construiu um conjunto de dados de pares de manchetes satíricas de contexto verdadeiro, no qual o contexto é construído recuperando e classificando processualmente histórias, eventos e informações verdadeiras relacionados às entidades que aparecem no título satírico original. Para isso, os autores estudaram o emprego de representações contextuais mais ricas através da capacidade de arquiteturas baseadas em transformadores, como o BERT (de *Bidirectional Encoder Representations from Transformers*)⁵ (DEVLIN et al., 2019), de gerar manchetes satíricas engraçadas e mostraram que tanto os modelos de linguagem quanto os modelos de sumarização podem ser ajustados para gerar as manchetes cômicas.

Ionescu e Chifu (2021) se concentraram na detecção da sátira de domínio cruzado (*cross-source*) a partir do FreSaDa (*French Satire Data*), um *corpus* de notícias coletadas de fontes de publicação reais e satíricas para o francês com base em uma abordagem superficial baseada em recursos de baixo nível, ou seja, caracteres n-gramas e um método profundo baseado em *embeddings* CamemBERT (MARTIN et al., 2020). Os autores compararam os dois métodos em duas configurações de classificação binária: (1) classificação de notícias verdadeiras completas *versus* sátira e (2) classificação de manchete verdadeiras *versus* sátira, observando que o modelo baseado em *embeddings* CamemBERT obtiveram melhores resultados em notícias verdadeiras completas, enquanto o modelo baseado em caracteres n-gramas alcançaram um desempenho superior na tarefa de detecção de sátira em manchetes de notícias é significativamente mais desafiadora, com a taxa de precisão máxima de 74,07%.

⁵ O BERT (Bidirectional Encoder Representations from Transformers) é um algoritmo de aprendizado profundo (*deep learning*) o PLN. É um modelo pré-treinado que é profundamente bidirecional e faz pouco uso de qualquer outra coisa além de um *corpus* de texto simples.

O modelo baseado em contexto utilizado na pesquisa de Horvitz, Do e Littman (2020, p. 42) pode capturar aspectos de “transformação de humor“ que inclui tabus e assuntos mais tensos. Além disso, o modelo parece conseguir imitar elementos do humor, como falsa analogia e usar relações incongruentes entre entidades e ideias. Os autores descobriram que os modelos das notícias apreenderam outros recursos utilizados nas notícias satíricas, por exemplo, o uso da justaposição de “um estudo” com uma observação científica, como a menção de eventos absurdos, mas contextualmente relevantes (“estudo descobre que a maioria dos americanos ainda está em derramamento de óleo”)⁶.

Em relação às pesquisas para notícias multilíngues, Guibon et al. (2019) compararam diferentes métodos para detecção de notícias falsas baseados em análise estatística de texto em um *corpus* de notícias falsas do inglês e do francês, assim como transcrições automáticas do YouTube em francês sobre a vacinação, devendo ser classificadas em notícias falsas, confiável ou sátira. Para isso, os autores utilizaram uma abordagem experimental, com foco no impacto da representação de dados para encontrar a melhor forma de classificar estes textos. Guibon et al. (2019) observaram que a semelhança e a detecção do domínio do texto, por si só, não conseguem lidar com toda a ambiguidade, sendo necessária a combinação de alguns métodos de classificação, como mineração de texto. Por fim, a comparação dos métodos de detecção de notícias mostrou que a combinação de métodos de representação e as *embeddings* fornecem resultados mais significativos.

3.2 AS CARACTERÍSTICAS LINGUÍSTICAS DE NOTÍCIAS SATÍRICAS

Burfoot e Baldwin (2009) foram os primeiros a retratarem uma classificação computacional de notícias satíricas. Os autores descreveram um método para filtrar artigos de notícias satíricas, contribuindo com a iniciativa de apresentar a detecção destas notícias à linguística computacional. O *corpus* utilizado em sua pesquisa consiste em 4.000 documentos de boletins de notícias e 233 artigos de notícias de sátiras, divididos em conjuntos fixos de treinamento (sendo 2505 notícias verdadeiras e 133 notícias satíricas) e teste (sendo 1495 notícias verdadeiras e 100 notícias satíricas).

Com base no *corpus*, Burfoot e Baldwin (2009) desenvolveram uma abordagem baseada em recursos lexicais e semânticos e *support vector machines* (SVMs) em características simples de *bag-of-words*. Os autores destacam três tipos de características destinados a identificar documentos de notícias satíricas, como destacado abaixo:

- **Características da manchete:** As notícias satíricas podem ser reconhecidas como tal apenas pelo título.
- **Profanidade:** Dificilmente uma notícia verdadeira irá incluir em sua redação uma linguagem ofensiva e as notícias satíricas podem se utilizar desse recurso como

⁶ Traduzido de “study finds majority of americans still in oil spill”

artifício humorístico.

- **Gírias:** Assim como o item anterior, as notícias verdadeiras procuram não utilizar gírias, podendo ocorrer com mais frequência em notícias satíricas.

Ainda em Burfoot e Baldwin (2009), segundo os autores, as abordagens lexicais não são suficientes para presumir que as notícias satíricas tendem a parodiar as notícias reais em tom, estilo e conteúdo, sendo necessária, portanto, uma abordagem semântica. Nesse contexto, os autores observaram que o *absurdo* é um recurso comum em notícias satíricas. Além disso, a validade semântica também foi introduzida usando o *Reconhecimento de Entidades Nomeadas* (REN)⁷, capaz de detectar se uma entidade nomeada está ou não fora do lugar, ou está sendo usada no contexto correto. Para isso, utilizou-se o reconhecimento das entidades nomeadas por meio da *validade*, que consiste na frequência relativa da combinação de participantes relatados na notícia, identificando-as em um documento e consultando sua ocorrência na Web para a conjunção dessas entidades. Os autores concluem que a tarefa de detecção de sátira em notícias é uma tarefa intrinsecamente semântica, uma vez que os artigos satíricos são reconhecíveis pela interpretação e comparação cruzada com o conhecimento de mundo dos leitores.

Outros trabalhos também compilaram *corpora* paralelos de notícias verdadeiras e satíricas, semelhante ao que é proposto nesta pesquisa. Em Rubin et al. (2016), a análise foi realizada a partir de duas frentes: i) a linguística – em que um linguista analisou o conteúdo de cada par (satírico e legítimo) com o objetivo de encontrar *insights* sobre as principais semelhanças e diferenças, assim como tendências no uso da linguagem e recursos retóricos; e ii) computacional – para o aprendizado de máquina. Para essa finalidade, foram coletados e analisados 360 artigos de notícias satíricas⁸, bem como suas contrapartes legítimas. O conjunto de dados é distribuído em 12 tópicos de notícias – com três tópicos distintos dentro de cada um dos quatro domínios (civismo, ciência, negócios e notícias “leves”).

Com base no *corpus*, analisou-se o conteúdo de cada par de notícias (legítimo *versus* satírico) a fim de encontrar semelhanças e diferenças entre elas, assim como predisposições no uso da língua e de dispositivos retóricos, destacando o uso do *absurdo* e do *humor* e a *complexidade da sentença*. A partir dessa análise linguística realizada, Rubin et al. (2016) notaram que o *absurdo* é um tipo de recurso que se mostrou útil na identificação de artigos satíricos. Descreve-se abaixo a proposta de um conjunto com cinco recursos linguísticos de notícias satíricas elaborados pelos autores:

⁷ Para o PLN, o Reconhecimento de Entidades Nomeadas (REN) é uma tarefa que procura identificar as Entidades Nomeadas de um texto, como nomes de pessoas, cidades e organizações, classificando-as em um conjunto pré-definido de categorias e remetendo a um referente específico.

⁸ Rubin et al. (2016) afirmam que para cada um dos 12 tópicos abaixo, eles coletaram cinco notícias satíricas canadenses (do *The Beaverton*) e cinco notícias americanas (do *The Onion*).

- **Absurdo:** É a introdução inesperada de novas entidades nomeadas (pessoas, lugares, organizações, entre outras) na última sentença de uma notícia satírica.
- **Humor:** Característica baseada nas premissas dos *scripts* opostos e a maximização da distância semântica entre duas declarações como método de identificação de uma piada.
- **Gramática:** Estabelece-se pelo conjunto de frequências de termos normalizados comparados com os dicionários *Linguistic Inquiry and Word Count* (LIWC).
- **Afeto Negativo:** Relaciona-se à frequência normalizada dos termos negativos presentes no texto das notícias.
- **Pontuação:** Relaciona-se à frequência normalizada das pontuações presentes no texto da notícia, como pontos, vírgula, ponto e vírgula, pontos de interrogação, pontos de exclamação e aspas.

Os autores ainda analisaram que a linha final de cada notícia satírica geralmente é uma piada, destacando ou um absurdo presente na sentença, ou introduzindo um novo elemento na história. Também foi observada por eles uma notável diferença sintática entre as notícias satíricas e verdadeiras em relação ao comprimento e a complexidade da sentença. Um teste com a medição de alguns índices da complexidade da sentença foi analisado no Capítulo 5 desta tese.

Yang, Mukherjee e Dragut (2017) observaram que as pistas satíricas são muitas vezes refletidas em certos parágrafos e não em toda a notícia e, assim, afirmam que considerar apenas uma análise no nível do documento para detectar a sátira pode ser uma tarefa limitada. Para a análise, foram coletadas notícias de 14 sites declaradamente satíricos. No total, soma-se mais 16 mil notícias satíricas e mais de 160 mil notícias verdadeiras de diversas fontes. Os autores aplicaram experimentos no *corpus* a partir de quatro categorias linguísticas para expor diferenças entre conteúdos genuínos e satíricos. Tais características foram computadas separadamente em parágrafo e documento, sendo classificadas como:

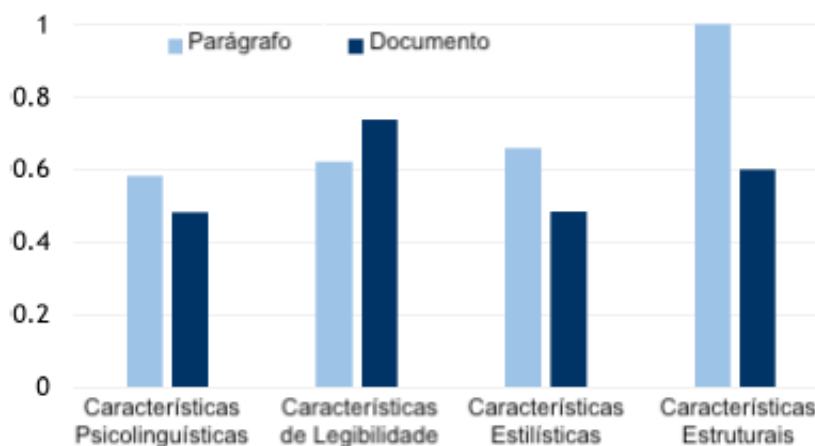
- **Características psicolinguísticas:** As diferenças psicológicas entre notícias satíricas e verdadeiras são úteis na detecção de notícias satíricas, pois, segundo os autores, os jornalistas profissionais tendem a expressar opiniões de forma mais conservadora, enquanto as satíricas podem apresentar uma linguagem mais agressiva para fins de entretenimento. Além disso, Yang, Mukherjee e Dragut (2017) observaram que enquanto as notícias verdadeiras estão focadas na clareza e na precisão da escrita, as notícias satíricas se relacionam às características emocionais. Desse modo, foi empregado o LIWC como dicionário psicolinguístico, sendo que cada categoria é considerada uma característica independente e calculada por sua frequência (contagem totais de cada categoria divididas pelo total de palavras).

- **Características estilísticas da escrita:** Refere-se à distribuição relativa de *tags* de *Part-of-Speech* (POS), o que para os autores, reflete a escrita informativa *versus* imaginativa, contribuindo na detecção de conteúdo enganoso. Cada *tag* é considerada uma característica independente e calculada por sua frequência.
- **Características de legibilidade:** Yang, Mukherjee e Dragut (2017) consideram que a legibilidade das notícias verdadeiras diferiria das notícias satíricas, uma vez que as primeiras são escritas por jornalistas e tendem a ser mais claras e objetivas, enquanto as últimas contém inúmeras orações para enriquecer a história que é inventada. Para isso, os autores utilizaram as seguintes métricas de legibilidade: *Flesch Reading Ease* (KINCAID et al., 1975), *Gunning Fog Index* (GUNNING, 1952), *Automated Readability Index* (SMITH; SENTER, 1967), *ColemanLiau Index* (COLEMAN; LIAU, 1975), além da contagem de sílabas por palavra.
- **Características estruturais:** Nesta categoria foram examinadas as seguintes características presentes no texto: contagem de palavras, contagem de palavras de \log^9 .

A partir das características linguísticas relacionadas acima, Yang, Mukherjee e Dragut (2017) comparam a importância dos quatro conjuntos de características no nível do parágrafo e no nível do documento. Primeiro foi relatada a importância escalonada dos quatro conjuntos de características linguísticas, calculando a média da importância das características individuais. Como mostra a Figura 12, a importância das características de nível do parágrafo é maior que os recursos do nível de documento, com exceção para o conjunto de legibilidade. Os autores enfatizam ser razoável o uso da legibilidade no nível de documento porque os recursos de legibilidade avaliam a facilidade de um determinado texto, que depende do conteúdo e da apresentação. Esta diferença entre níveis de análise também é observada no Capítulo 5, pois enquanto a análise manual de recursos linguísticos ocorre no nível da sentença, a análise da complexidade textual das notícias pelo NILC-Metrix (LEAL, 2021) ocorre no nível do documento.

⁹ Refere-se a um registro (log) de contagem de palavras, como se, por algum motivo, a ferramenta (ou o método, ou a técnica) tivesse que fazer várias contagens de palavras e registrando tais contagens.

Figura 12 – Comparação da importância dos quatro conjuntos de características no nível do parágrafo e no nível do documento



Fonte: Adaptado de Yang, Mukherjee e Dragut (2017, p. 1985).

A avaliação sugere que em relação ao nível de documento, o número de palavras, letras maiúsculas e pontuações nas notícias verdadeiras são maiores que nas notícias satíricas, mas quando são analisados no nível do parágrafo, esses recursos nas notícias verdadeiras são menores que nas notícias satíricas. Os autores também apuraram que embora as notícias satíricas sejam mais curtas do que as notícias verdadeiras no nível do documento, quando se refere ao nível do parágrafo, as notícias satíricas geralmente contêm parágrafos mais complexos do que as notícias verdadeiras. Finalmente, Yang, Mukherjee e Dragut (2017) concluem que na análise das características individuais revela que a escrita de notícias satíricas tende a ser emocional e imaginativa.

No trabalho de Barbieri, Ronzano e Saggion (2015), os autores analisaram perfis no Twitter, sendo dois perfis satíricos e dois perfis de jornais não satíricos em cada dos três idiomas explorados – inglês, espanhol e italiano. Logo, obteve-se um conjunto de dados balanceados com 2.766 *tweets* de cada conta selecionada, resultando o total de 33.192 *tweets* (sendo 16.596 de notícias satíricas e 16.596 de notícias não satíricas). A partir do *corpus* coletados, cada *tweet* foi caracterizado em sete classes de características:

- **Baseado em palavras:** Recurso calculado a partir de cinco recursos baseados em palavras: lemas, bigramas (combinação de dois lemas em uma sequência) e salto 1, 2 e 3 grama.
- **Frequência:** Foram extraídos três tipos de características de frequência: frequência da palavra mais rara (frequência da palavra mais rara incluída no *tweet*), média de frequência (a média aritmética de todas as frequências das palavras no *tweet*) e intervalo de frequência (a diferença entre os dois recursos anteriores).

- **Sinônimos:** Consideram-se as frequências dos sinônimos de cada palavra no *tweet* (para cada idioma seu próprio *corpora* de frequência), conforme recuperados do WordNet. Em sequência, foram calculados o maior e o menor número de sinônimos com frequência maior que a presente no *tweet*, o número médio de sinônimos com frequência maior/menor que a frequência da palavra relacionada presente no *tweet*. Além disso, determinou-se também o maior/menor número de sinônimos e o número médio de sinônimos das palavras com frequência maior/menor que a presente no *tweet*. Por fim, os autores calcularam o conjunto de recursos de Sinônimos considerando todas as palavras do *tweet* juntas e apenas as palavras pertencentes a cada uma das quatro *part-of-speech*.
- **Ambiguidade:** Para modelar a ambiguidade das palavras nos *tweets*, os autores utilizaram os *synsets* da WordNet associados a cada palavra. A hipótese é que se uma palavra inclui vários significados(/*synsets*), é provável em que seja utilizada de forma ambígua. Desse modo, para cada *tweet*, eles calcularam o número máximo de *synsets* associados a uma única palavra, o número médio de *synsets* de todas as palavras e o intervalo de *synset* que é a diferença entre os dois recursos anteriores. Determinaram, por fim, o valor desses recursos incluindo todas as palavras de um *tweet*, bem como considerando separadamente apenas nomes, verbos, adjetivos ou advérbios.
- **Part-of-Speech:** Os recursos incluídos no grupo *Part-of-Speech* (POS) são pensados para capturar a estrutura sintática dos *tweets*. As características deste grupo são oito (verbos, nomes, adjetivos, advérbios, interjeições, determinantes e pronomes) e cada uma delas conta o número de ocorrências de palavras caracterizadas por um determinado POS.
- **Sentimentos:** Os sentimentos das palavras nos *tweets* são importantes por dois motivos: i) para detectar o sentimento (por exemplo, se os *tweets* compreenderem termos positivos ou negativos) e para apreender o inesperado criado por uma palavra negativa em um contexto positivo ou vice-versa. Com base em um léxico de sentimento, os autores computaram o número de palavras positivas/negativas, a soma das intensidades das pontuações positivas/negativas das palavras, a média da pontuação positiva/negativa das palavras, a maior pontuação positiva/negativa pontuação, a diferença entre a maior pontuação positiva/negativa e a média positiva/negativa. Também mediram a proporção das palavras com polaridade diferente de zero, para detectar subjetividade no *tweet*. Para esta tarefa, foram considerados apenas os nomes, verbos, adjetivos e advérbios.
- **Pontuação:** Esse recurso foi pensado para capturar o estilo de pontuação dos *tweets* satíricos. Considera-se nessa característica o conjunto do número de um sinal de

pontuação específico, que inclui: “.”, “!”, “?”, “\$”, “%”, “”, “+”, “-”, “=”. Também foi computado os números de caracteres maiúsculos e minúsculos e o comprimento do *tweet*.

A abordagem proposta por Barbieri, Ronzano e Saggion (2015) evita o uso de recursos como *bag-of-words*, contando apenas com recursos intrínsecos da palavra, uma vez que eles visam detectar características internas das palavras. Ao testar a abordagem em *tweets* em inglês, espanhol e italiano, os autores observaram que ao treinar em italiano, o sistema não é capaz de detectar sátiras em inglês e espanhol, mas ao testar em italiano e treinar em outros idiomas, os resultados são melhores. Barbieri, Ronzano e Saggion (2015) acreditam que a interpretação pode ser que a sátira italiana é menos intrincada, fácil de detectar, mas incapaz de reconhecer outro tipo de sátira. Além disso, o modelo de palavra intrínseca quando treinado em espanhol é capaz de detectar a sátira italiana com uma precisão de 0,695, o que segundo eles, é um resultado muito interessante considerando a complexidade da tarefa. Por fim, afirmam que a precisão de seu modelo é promissora (0,767), pois neste conjunto de dados o ruído é muito alto, pois, são 22.228 *tweets* em três idiomas e tópicos diferentes.

Cabe citar que característica como as *baseado em palavras* busca capturar padrões, como a frequência e o sentimentos de cada palavra no texto, uma vez que estes aspectos são intrínsecos às palavras e ao texto. Desse modo, os autores pontuam que esses tipos de características não dependem de padrões lexicais de uma língua específica e podem, portanto, ser usadas em vários idiomas.

No trabalho de Levi et al. (2019), é abordado o desafio de classificar automaticamente notícias falsas *versus* notícias satíricas, baseando-se na hipótese que determinadas nuances textuais entre estes tipos de notícias podem ser identificadas usando pistas semânticas e linguísticas. Para a tarefa, os autores treinaram um método de aprendizado de máquina usando:

- **Representação semântica com BERT:** os autores utilizaram o BERT para estudar as nuances semânticas entre notícias falsas e satíricas.
- **Análise linguística com Coh-Matrix:** os autores partiram da hipótese que métricas de coerência textual são úteis para capturar aspectos semelhantes de relação semântica entre diferentes sentenças de uma notícia.

Levi et al. (2019) avaliaram o conjunto do método com base no conjunto de dados de notícias falsas e notícias satíricas. A avaliação dos métodos apontou para a existência de diferenças semânticas e linguísticas entre as notícias falsas e notícias satíricas. Tanto a representação semântica com o BERT (DEVLIN et al., 2019) quanto a análise linguística com o Coh-Matrix (MCNAMARA et al., 2010) alcançaram um desempenho significativamente melhor do que o método em *baselines*. Uma das nuances observadas

pelos autores é que as notícias satíricas são mais sofisticadas ou menos fáceis de ler em relação às notícias falsas. Nesta pesquisa, não são consideradas as notícias falsas, mas como é apresentado no Capítulo 5, as notícias satíricas são mais fáceis de ler quando compradas às notícias verdadeiras.

Diferentemente da proposta desta tese, que tem em vista encontrar mecanismos linguísticos presentes na estrutura textual das notícias satíricas, para o português europeu, Carvalho et al. (2020) investigaram dispositivos linguísticos e retóricos de ironia implícitos em manchetes satíricas com o propósito de automatização desse tipo de conhecimento. O estudo baseia-se em um *corpus* de manchetes irônicas, extraídas do *O Inimigo Público* (um semanário satírico do jornal português *O Público*), do qual se constitui de 2750 manchetes anotadas, por cinco anotadores, com informações sobre os dispositivos retóricos subjacentes a elas, seguindo diretrizes detalhadas e desenvolvidas especificamente para essa tarefa. O estudo identificou categorias diferentes e algumas delas se enquadram na definição de figuras de linguagem clássicas:

- Antítese, comparação, hipérbole, metáfora, paradoxo, paralelismo, repetição e vulgarismo.

Além disso, os autores também encontraram o uso de propriedades sintáticas e semânticas:

- Uso expressivo de adjetivo, diminutivo, conectivos de exclusão (como *exclusive*, *menos*, *exceto*, *fora*, *salvo*, *senão*, *sequer*, *apenas*), ruptura distributiva e idiomática e contraste fora do domínio (combinações incomuns de substantivos e entidades nomeadas).

Com base nos dados levantados, os autores descobriram que a incongruência presentes em textos irônicos envolve, muitas vezes, contrastes fora do domínio, podendo ser apreendido com sucesso no contraste estatístico de co-ocorrência de entidades nomeadas dentro de um título específico em relação àqueles historicamente relacionados semanticamente em repositórios de dados. Carvalho et al. (2020) também descobriram que, diferente da ironia verbal, características de sentimento, como polaridade, não são eficazes na identificação da ironia situacional¹⁰. Por fim, afirmam que a prevalência de adjetivos nas manchetes pode sinalizar uma intenção irônica, mas que as características utilizadas para identificação de incongruência em textos sarcásticos não são relevantes para detectar a ironia situacional em notícias satíricas do português.

Sobre o português do Brasil, Wick-Pedro et al. (2020), investigaram como as notícias satíricas são reconhecidas e compartilhadas pelos usuários do Twitter e quais características

¹⁰ No trabalho de Carvalho et al. (2020), os autores afirmam que a função da ironia situacional é enfatizar eventos (reais ou fictícios) que evocam imagens peculiares e inesperadas, que costumam criar um efeito cômico no público.

linguísticas mais relevantes nos comentários dos *tweets* de notícias satíricas. Os dados foram coletados automaticamente pelo *GATE Cloud*¹¹, somando aproximadamente 36 mil *tweets* e *retweets* relacionados ao site *Sensacionalista*. Grande parte das notícias tem como assunto principal conteúdo político e, sobretudo, como alvo principal o atual Presidente Jair Bolsonaro. Do total de *tweets* coletados, uma parcela de aproximadamente 18 mil *tweets*, foram separados em: i) comentários e *replies*: além de compartilhar, usuário faz um comentário e/ou emite uma opinião sobre a notícia; ii) *retweets*: o usuário apenas compartilha a notícia; iii) links: o usuário utiliza *gifs* ou imagens e iv) postagens do Sensacionalista: são postagem do site *Sensacionalista* no Twitter.

Após uma atenta e minuciosa leitura, os *tweets* classificados em comentários e *replies* foram analisados e classificados em 5 categorias de acordo com a interpretação e intenção do usuário em relação à notícia postada (WICK-PEDRO et al., 2020):

- **Categoria 1:** o usuário vai emitir sua opinião, genericamente, sobre o conteúdo da notícia ou sobre assuntos relacionados à notícia postada.

(6) O curioso é que o pão francês na verdade não é francês.

- **Categoria 2:** comentários em que a sátira é compreendida pelo usuário.

(7) Kkkkkkkk essa foi muito engraçada

- **Categoria 3:** comentários tóxicos sobre o assunto/alvo da notícia.

(8) Tá feliz com o monstro que você criou? Justifique-se.

- **Categoria 4:** comentários positivos sobre o assunto/alvo da notícia.

(9) Sérgio Moro é sensacional, incomodou geral.

- **Categoria 5:**¹² comentários duvidosos.

(10) Isso aí. Vamos arrumar o Brasil. Um desafio!!!

Em sequência à categorização dos comentários, foram considerados alguns aspectos linguísticos de cada categoria, como *hashtags*, *emoticons*, risadas, adjetivos, aspas, pontuação e diminutivo. Por meio dessa análise, Wick-Pedro et al. (2020) observaram que o uso dos adjetivos é necessário para o direcionamento da opinião (positiva ou negativa) do usuário em relação às notícias compartilhadas.

¹¹ Disponível em: <<https://cloud.gate.ac.uk/>>

¹² Nesta categoria um comentário pode ser entendido como irônico, mas também há dúvida sobre a compreensão da sátira pelo usuário.

Os autores concluem que embora a maioria dos usuários se utilizam das notícias satíricas para poder comentar não só sobre a notícia, mas sim sobre todo o contexto em que a notícia está inserida. Porém, há ainda uma grande interação dos usuários que compreendem o efeito satírico e de humor presentes nas notícias do *Sensacionalista* (WICK-PEDRO et al., 2020).

Sobre a detecção de notícias em diferentes línguas, Abonizio et al. (2020) avaliaram recursos textuais que não estão atrelados a uma linguagem específica na descrição de dados textuais para detecção de notícias. Para essa avaliação, os autores utilizaram um *corpora* multilíngue de notícias para o inglês americano, espanhol e português brasileiro, explorando a complexidade, estilometria e recursos psicológicos presentes no texto. O *pipeline*¹³ da pesquisa é composto por três etapas: pré-processamento, extração de características e classificação, abrangendo o tratamento das notícias brutas até o resultado da detecção.

Na etapa de pré-processamento foram eliminadas as características indesejadas durante a coleta de dos dados. Nesta fase, tanto os caracteres não textuais (por exemplo, *emoticons* e caracteres especiais), como os metadados dos sites foram removidos, deixando apenas os dados textuais relacionados a notícias. Nota-se que esses tipos de elementos não foram extraídos nesta atual pesquisa, por acreditar ser um elemento não só textual, mas também linguístico e que pode caracterizar uma notícia satírica. Ainda foram filtrados pequenos textos para evitar notícias muito curtas, além de criar uma homogeneidade nos comprimentos dos textos analisado. Por fim, foram removidos espaços em branco extras, assim como outros ruídos relacionados à coleta e processamento que não estão relacionados ao conteúdo da notícia.

Em relação à etapa de extração de características, Abonizio et al. (2020) propuseram características independentes da linguagem, concentrando-se na captura de alto nível, uma vez que o mesmo conjunto de características possa ser usado em domínios multilíngues. Desse modo, os autores extraíram características de três categorias: complexidade, estilística e psicologia. As características baseadas na complexidade visam capturar a **complexidade** das notícias, tanto no nível da sentença quanto da palavra. Para isso, foram usadas métricas que calcularam o **tamanho médio das palavras**, a **contagem de palavras por sentença** e o *Type-Token Ratio* (TTR). As características estilísticas, por sua vez, utilizaram técnicas do PLN para extrair informações gramaticais de cada documento. Assim, utilizou-se de *tagger part-of-speech* (POS) para rastrear a frequência de classes gramaticais. Outro recurso interessante utilizado pelos autores foi o *Out-Of-Vocabulary* (OOV), um método que utiliza um dicionário de palavras para um determinado idioma e conta o total de palavras que não são encontradas nesse conjunto e suas frequências no texto, tendo como objetivo capturar neologismos, gírias ou outros tipos de palavras inusitadas.

¹³ Um pipeline consiste em uma cadeia de elementos de processamento – processos, *threads*, funções – dispostos de modo que a saída de cada elemento seja a entrada do próximo. Este método aumenta o número de instruções executadas simultaneamente e a taxa de instruções iniciadas e terminadas por unidade de tempo.

Enfim, as características psicológicas estão relacionadas aos processos cognitivos do texto. Desse modo, foi avaliada a polaridade do texto, medindo a positividade ou negatividade de um documento.

Na etapa de automatização as características foram exploradas em modelos de aprendizado de máquina, o que indicou que os recursos puramente estilométricos, como a diversidade gramatical dos *tags* POS, a proporção de entidades nomeadas para o tamanho do texto, a proporção de aspas para o tamanho do texto e a frequência de palavras OOV, conseguiram melhorar a previsão resultados. Abonizio et al. (2020) concluem que o padrão compartilhado entre as linguagens estudadas sugere a existência de um comportamento subjacente entre as diferentes linguagens, que pode suportar a detecção de fake news em vários idiomas além dos já abordados.

Além dos trabalhos que abordaram o LIWC como uma ferramenta para extração de características linguísticas para a detecção de notícias satíricas (RUBIN et al., 2016; YANG; MUKHERJEE; DRAGUT, 2017), Salas-Zárate et al. (2017) propuseram um método que emprega uma ampla variedade de recursos psicolinguísticos para a detecção de textos satíricos e não satíricos para o espanhol. Assim, os autores analisaram características psicolinguísticas para captar não só o conteúdo, mas também o estilo da mensagem a partir do LIWC visando determinar quais são os aspectos mais discriminantes para detecção de sátira, uma vez que, segundo eles, a aplicação do LIWC foi bem sucedida no que diz respeito à linguagem figurada. O estudo foi realizado usando um *corpus* composto por oito contas do Twitter, quatro são satíricas (duas da Espanha e duas do México) e quatro não são satíricas (duas da Espanha e duas do México), diferentemente do que foi realizado por Barbieri, Ronzano e Saggion (2015), que só considerou a variante peninsular. No total, o *corpus* conta 10.000 *tweets* satíricos (5.000 da Espanha e 5.000 do México) e 10.000 *tweets* não satíricos (5.000 da Espanha e 5.000 do México).

O método de detecção de sátira proposto por Salas-Zárate et al. (2017) é dividido em três etapas principais: (i) pré-processamento de texto, que envolve a limpeza e correção do *corpus*; (ii) extração de características, que consiste na extração das características psicológicas e linguísticas por meio do LIWC e, por fim, (iii) treinamento do algoritmo de aprendizado de máquina, que consiste no treinamento de algoritmos de classificação de aprendizado de máquina. Em relação à extração das características psicológicas e linguísticas, ao contrário de trabalhos anteriores, que analisaram apenas algumas características (como emoções negativas, negações, quantificadores, certeza, tentativa, exclusão, inclusão, discrepância, causalidade, passado, presente e sinais de pontuação) Salas-Zárate et al. (2017) examinaram as 72 categorias disponíveis no LIWC com o propósito de detectar quais dos recursos para linguagem figurada mencionados na literatura e outros recursos do LIWC podem contribuir para a detecção de sátira. Essas categorias são classificadas adicionalmente em cinco conjuntos principais:

- **Processos linguísticos:** Relaciona-se a informações gramaticais como o total de pronomes, artigos, negações, contagem de palavras e verbos auxiliares, entre outros;
- **Processos psicológicos:** Considera emoções positivas, emoções negativas, processos sociais e processos cognitivos, entre outros;
- **Preocupações pessoais:** Conjunto composto por categorias de palavras relacionadas a preocupações pessoais intrínsecas à condição humana;
- **Categorias faladas:** Estabelece certas dimensões da linguagem falada;
- **Pontuação:** Considera doze categorias de pontuação – pontos, vírgulas, dois pontos, ponto e vírgula, pontos de interrogação, pontos de exclamação, travessões, aspas, apóstrofos, parênteses, outra pontuação.

Os autores utilizaram o WEKA¹⁴ (BOUCKAERT et al., 2010) para treinar os algoritmos de classificação de aprendizado de máquina. Esse trabalho foi baseado em algoritmos de classificação de aprendizado de máquina conhecidos como classificadores, que permitem a criação de modelos de acordo com os dados e finalidade da análise. Nesta fase, procurou-se um modelo baseado na análise das instâncias, representado por meio de regras de classificação, árvores de decisão ou fórmulas matemáticas. Assim, para distinguir textos satíricos e não satíricos, foi realizado um conjunto de experimentos com o objetivo de medir a eficácia da abordagem proposta para a detecção de *tweets* satíricos.

Como resultado, Salas-Zárate et al. (2017) observaram que as características linguísticas, as características emocionais e os sinais de pontuação são importantes na tentativa de detecção da sátira. Além disso, os resultados demonstram que os *tweets* satíricos contêm mais advérbios, quantificadores e verbos no tempo presente quando comparado aos *tweets* não satíricos. Em relação ao processo psicológico, cinco características foram preditores significativos: processo social, processo afetivo, emoções positivas, processo cognitivo e certeza. A respeito dos sinais de pontuação, observou-se que dois pontos, pontos de exclamação, apóstrofos e aspas foram preditores significativos para a detecção de *tweets* satíricos. Finalmente, concluem que os *tweets* satíricos contêm palavras mais positivas para alterar o significado de uma afirmação negativa, ou seja, um problema social, abusos ou qualquer situação ultrajante que o usuário deseje divulgar. Também atribuem a alta taxa do processo cognitivo ao fato de que a sátira é mais complexa e difícil de entender do que a linguagem literal.

Por fim, García-Díaz e Valencia-García (2022) avaliaram textos satíricos a partir de várias arquiteturas e conjuntos de recursos de aprendizado profundo e, assim, computaram quais são os recursos fortes para identificação automática de sátira. Para isso, criaram

¹⁴ O WEKA é uma coleção de algoritmos de aprendizado de máquina que podem ser usados para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização.

o SatiCorpus 2021, um *corpus* composto por *tweets* satíricos e não satíricos compilados principalmente de sites de notícias. Após o pré-processamento e limpeza dos dados, os autores dividiram o *corpus* em treinamento, validação e teste numa proporção de 60-20-20. Em seguida, conduziram a extração de recursos para obter os recursos linguísticos e os recursos baseados em incorporação e, finalmente, os melhores modelos para cada conjunto de recursos são avaliados com o conjunto de dados de teste.

As características linguísticas utilizadas por García-Díaz e Valencia-García (2022) foram extraídas pela ferramenta a UMUTextStats (GARCÍA-DÍAZ; CÁNOVAS-GARCÍA; VALENCIA-GARCÍA, 2020; GARCÍA-DÍAZ et al., 2021), inspirada no LIWC (TAUSCZIK; PENNEBAKER, 2009), mas pensada para a língua espanhola. O UMUTextStats consegue capturar um total de 365 recursos linguísticos diferentes, nos quais foram organizados nas 10 categorias abaixo:

- **Correção e estilo:** Esta categoria está relacionada à gramaticalidade do texto, como (1) erros ortográficos, relacionados ao número de palavras com erros ortográficos ou ao mau uso da acentuação do espanhol; (2) erros estilísticos, relacionados à presença de frases que começam com números ou com a mesma palavra e (3) desempenho, relacionado à busca de palavras duplicadas, uso de ponto final após ponto exclamação ou interrogação e expressões redundantes.
- **Fonética:** Está relacionada ao alongamento expressivo de algumas letras com o propósito de enfatizar algo ou o que está sendo dito, considerando apenas a repetição de três ou mais letras e, no caso das vogais, se elas têm ou não acento.
- **Morfossintaxe:** Considera-se nesta categoria as características que representam como as palavras e sentenças são compostas. Especificamente, esta categoria captura o número e o gênero da palavra; afixos, para observar uma variedade refinada de sufixos (nominais, adjetivizadores, verbalizadores, adverbializadores, aumentativos ou diminutivos) e prefixos. Esta categoria também está relacionada às classes de palavras.
- **Semântica:** Esta categoria reúne quatro características semânticas. São elas: (1) onomatopeia, refere-se às palavras que compõem em relação ao som que produzem; (2) eufemismo, recursos linguísticos utilizados para dizer de modo mais suave o que poderia ser considerado muito grosseiro em alguns contextos; (3) disfemismo, palavras vulgares utilizadas para substituir outras mais neutras e (4) sinédoque, um tipo de tropo literário usado para representar uma parte como um todo, por exemplo, quando se diz “tenho quatro bocas para alimentar”, na verdade, “quatro bocas” se refere a quatro crianças.
- **Pragmática:** Relaciona-se à presença de dispositivos de linguagem figurada, como eufemismos, hipérboles, expressões idiomáticas, perguntas retóricas, ironia verbal,

metáforas, símiles, entre outros. Esta categoria comporta também recursos linguísticos para observar como as sentenças são conectadas por meio de marcadores de discurso.

- **Estilometria:** Está relacionada à contagem de (1) variedade de símbolos e pontuação no texto; (2) estatísticas do *corpus*, como *Token-Type Ratio* (TTR) e (3) outras métricas relacionadas ao número de palavras, sílabas ou sentenças.
- **Lexical:** Considera-se nesta categoria os tópicos abordados no texto. Para isso, os autores analisaram desde tópicos abstratos, como pensamento analítico, conquista, amizade, religião, certeza, entre outros, até tópicos gerais, como locais, organizações, animais, roupas, alimentações e profissões.
- **Processos psicolinguísticos:** Esta categoria está relacionada ao léxico e emojis relacionados a sentimentos (positivos e negativos) e emoções (raiva, tristeza, ansiedade).
- **Registro:** Esta categoria reúne característica de como as pessoas usam a linguagem para se comunicar, como a presença de linguagem informal ou culta. Também foram capturados tópicos relacionados ao discurso ofensivo.
- **Jargão de mídia social:** Nesta categoria, consideram-se características a pistas que revelam o domínio do falante no jargão de mídia social, pois pode ser uma terminologia específica usada nas mídias sociais ou o uso de mecanismos como hiperlinks, menções ou emojis.

Além das categorias linguísticas, García-Díaz e Valencia-García (2022) também avaliaram as *word embeddings* não contextuais (WE). Os autores também avaliaram os modelos de *embeddings* de sentenças não contextuais (SE) e *embeddings* contextuais de sentenças (BF). Como resultado, notou-se que a identificação da sátira por meio de características linguísticas é mais confiável com o uso de pistas estilométricas do que outras pistas como semântica, pragmática ou por sentimento. Além disso, García-Díaz e Valencia-García (2022) observaram não haver uma categoria linguística que se destaque das outras e que elas são complementares com todos os tipos de *embeddings*, independentemente de serem baseadas em palavras ou sentenças, ou de serem contextuais, ou não contextuais.



Os trabalhos apresentados neste capítulo mostram que apesar de existir uma variedade para alguns idiomas, a grande parte das pesquisas aqui abordadas referentes à detecção de sátira são para o inglês. Também foram abordados trabalhos que tratam a sátira a partir de um viés computacional, propondo sistemas de classificação e/ou avaliação

automática de notícias satíricas. Outras abordagens também propuseram características extraídas sobretudo de abordagens voltadas ao aprendizado de máquina que orientam uma análise linguística. Assim, diferente das propostas anteriores, neste estudo busca-se evidenciar uma análise linguística mais aprofundada, além de ser auxiliada com ferramentas linguístico-computacionais.

O Quadro 3 apresenta a relação dos autores, o idioma tratado, as principais características linguísticas extraídas e as características, destacadas em negrito, que também foram abordadas nesta pesquisa. Dos trabalhos relacionados neste capítulo, serão consideradas as características descritivas, como contagem de sílabas por palavras e contagem de palavras usadas por Yang, Mukherjee e Dragut (2017), Abonizio et al. (2020) e García-Díaz e Valencia-García (2022), além do TTR, empregado por Abonizio et al. (2020) e García-Díaz e Valencia-García (2022). Ainda são consideradas as palavras de baixo calão, como profanidade e gírias, usadas por Burfoot e Baldwin (2009). Nota-se também que a pontuação é uma característica relevante nos trabalhos de Barbieri, Ronzano e Saggion (2015) e Salas-Zárate et al. (2017) e será incluída nos resultados da análise morfossintática de *part-of-speech* (POS), que também foram utilizadas por Barbieri, Ronzano e Saggion (2015), Yang, Mukherjee e Dragut (2017), Abonizio et al. (2020) e García-Díaz e Valencia-García (2022). Vale ressaltar que embora o BERT não tenha sido empregado nesta tese, como em Levi et al. (2019), o NILC-Matrix – uma versão do Coh-Matrix para o português do Brasil – foi usado para calcular as métricas de inteligibilidade textual.

Destaca-se ainda o trabalho de Carvalho et al. (2020) por descrever textos em português na variante europeia e pela descrição de manchetes satíricas com base em características retórico-estilísticas, como hipérbole, metáfora, ruptura distributiva e contraste fora de domínio, que também foram consideradas na análise realizada no Capítulo 5. Além disso, ainda foram aplicadas as características repetição e vulgarismo. É importante ressaltar que apesar do trabalho de Carvalho et al. (2020) ser para o português, assim como o descrito neste doutorado, ele considera apenas manchetes de notícias, focando somente na ironia situacional.

Quadro 3 – Principais características dos trabalhos relacionados

Autores	Idioma	Características utilizadas
Burfoot e Baldwin (2009)	Inglês	Características da manchete; profanidade; gírias.
Barbieri, Ronzano e Saggion (2015)	Multilíngue	Baseado em palavras; frequência; sinônimos; ambiguidade; <i>part-of-speech</i>; sentimentos; pontuação.
Rubin et al. (2016)	Inglês	Absurdo; humor; gramática; afeto negativo; pontuação.
Yang, Mukherjee e Dragut (2017)	Inglês	Características psicolinguísticas a partir do LIWC; <i>part-of-speech</i>; características de legibilidade e contagem de sílabas por palavras; contagem de palavras.
Salas-Zárate et al. (2017)	Espanhol	Contagem de classes gramaticais; processos psicológicos com base no LIWC; preocupações pessoais; categorias faladas; pontuação.
Levi et al. (2019)	Inglês	Representação semântica com BERT; análise linguística com Coh-Metrix.
Carvalho et al. (2020)	Português Europeu	Antítese; comparação; hipérbole; metáfora; paradoxo; paralelismo; repetição; vulgarismo; uso expressivo de adjetivo, diminutivo, conectivos de exclusão; ruptura distributiva e idiomática e contraste fora do domínio.
Abonizio et al. (2020)	Multilíngue	Métricas que calculam o tamanho médio das palavras, contagem de palavras por sentença, TTR; <i>part-of-speech</i>, <i>Out-Of-Vocabulary</i>.
García-Díaz e Valencia-García (2022)	Espanhol	Correção e estilo; fonética; morfofossintaxe; onomatopeia; eufemismo; disfemismo; sinédoque; presença de dispositivos de linguagem figurada; variação de símbolos e pontuações; TTR; métricas relacionadas ao número de palavras, sílabas ou sentenças; lexical; processos psicolinguísticos; registro; jargão de mídia social.

Fonte: Elaborado pela autora.

Por fim, o próximo capítulo apresenta o *corpus* e *subcorpus* construído para a análise linguística (cf. Capítulo 5) e suas principais características, bem como as ferramentas linguístico-computacionais utilizadas e o processo de extração de características linguísticas nas notícias satíricas.

4 MÉTODOS E MATERIAIS

Neste capítulo apresentam-se a metodologia e os materiais utilizados na pesquisa. Para alcançar os objetivos (cf. Seção 1.1), orientou-se metodologicamente esta pesquisa em quatro etapas principais:

Etapa 1 - Seleção do *corpus* da pesquisa: consiste na criação de um *corpus* de notícias satíricas para o português do Brasil (cf. Seção 4.1). Assim, os textos obtidos neste trabalho são provenientes de um *corpus* composto por notícias satíricas extraídas automaticamente do site do Sensacionalista, um portal de notícias satíricas para o português brasileiro.

Etapa 2 - Criação do *subcorpus* de análise: fundamenta-se na elaboração de um *subcorpus* constituído de notícias verdadeiras e suas referentes satíricas para a análise comparativa de cada tipo de texto. A construção desse tipo de *corpus* é necessária para procedimentos comparativos de construções linguísticas entre os textos reais e os satíricos (cf. Seção 4.1.3). O modelo desse *subcorpus* é semelhante ao Fake.Br Corpus apresentado por Monteiro et al. (2018), que consiste em notícias verdadeiras e falsas alinhadas manualmente para o português do Brasil.

Etapa 3 - Escolha de ferramentas linguístico-computacionais: baseia-se em apresentar quais as principais ferramentas utilizadas para auxiliar o processamento e a extração de dados linguísticos e estatísticos para a análise e descrição das notícias. Para a anotação morfosintática do *subcorpus* foi utilizado o *parser* PALAVRAS (BICK, 2000). Para o cálculo de medidas de inteligibilidade textual, como índices para avaliar coesão, coerência e dificuldade de compreensão de um texto, utilizando diversos níveis de análise linguística (como descritiva, lexical, sintática, semântica, coesiva e coerente, entre outras), utilizou-se o NILC-Matrix (LEAL, 2021). Para a análise textual automática do texto com base em recurso lexical foi utilizada a versão para o português do Brasil (BALAGE FILHO; PARDO; ALUÍSIO, 2013) do *Linguistic Inquiry and Word Count* (LIWC) (PENNEBAKER; FRANCIS; BOOTH, 2001).

Etapa 4 - Análise dos dados: apoiada nos dados extraídos na etapa anterior e com o *subcorpus* da Etapa 2, foi realizada uma análise linguística, (cf. no Capítulo 5), com o propósito de encontrar e descrever mecanismos linguísticos que podem ser utilizados como características linguísticas no reconhecimento automático de notícias satíricas.

4.1 CORPUS DA PESQUISA

Nesta seção é descrito o processo de seleção e construção do SatiriCorpus.Br¹, o *corpus* utilizado na pesquisa, assim como suas principais características e a construção do *subcorpus* para a análise linguística.

4.1.1 Seleção do *corpus*

Considerando que o leitor esteja atento ao humor presente no texto da notícia sátira, dificilmente a compreenderá como uma verdade (RUBIN; CHEN; CONROY, 2015). No entanto, para evitar que o público mais desatento ou aquele que não conhece o conteúdo dos jornais satíricos, estas páginas de sátira noticiosa procuram sempre alertar ao seu leitor sobre a informação humorística presente em suas notícias. Como mostra a Figura 13, o próprio Sensacionalista se autodenomina um “*jornal isento de verdade*”.

Figura 13 – Cabeçalho do site do Sensacionalista



Fonte: Página do Sensacionalista.²

Apesar da existência de outros portais (Piauí Herald³ e O Bairrista⁴) com a proposta de um jornalismo satírico, a escolha pelo Sensacionalista se dá por ele ser o maior expoente desse tipo de conteúdo no Brasil. Criado em 2009, o site do Sensacionalista possui quatro redatores: os jornalistas Martha Mendonça, Nelito Fernandes, Marcelo Zorzanelli e o historiador Leonardo Lanna. Em 2016, o site obteve mais de 11 milhões de acessos mensais (PESSOA, 2016) e, atualmente, seu perfil do Twitter⁵ possui 2,3 milhões de seguidores e 3 milhões no Facebook⁶. A Figura 14 exemplifica a atual página do Sensacionalista.

¹ Disponível em: <https://opencor.gitlab.io/corpora/wick_pedro21satiricorpus/>

² Cabeçalho do antigo site do Sensacionalista antes da desativação. Disponível em: <<https://www.sensacionalista.com.br/>>. Acesso em: 25 de out. 2020.

³ Disponível em: <<https://piaui.folha.uol.com.br/herald/>>

⁴ Disponível em: <https://www.twitter.com/o_bairrista>

⁵ Disponível em: <<https://twitter.com/sensacionalista>>

⁶ Disponível em: <<https://pt-br.facebook.com/sensacionalista/>>

Figura 14 – Exemplo do atual site do Sensacionalista

The screenshot shows the homepage of the Sensacionalista website. On the left, there is a logo with a yellow circle and a stylized 'S' and 'A'. Below it, the text 'SENSACIONALISTA' is displayed, along with social media icons for Facebook, Twitter, and Instagram. A search bar is present with the placeholder text 'Buscar neste blog'. Below the search bar, it says 'O jornal isento de verdade. Fundado em 2009.' and 'Quem escreve' followed by the Sensacionalista logo and a link 'Ver todos os blogs'.

The main content area features a headline under the 'HUMOR' category: 'Bolsonaro faz exames no cérebro para descobrir causa da obstrução fecal'. The author is 'Por Sensacionalista' and the date is '03/01/2022 - 11:17'. Below the headline is a photograph of Jair Bolsonaro lying in a hospital bed, looking unwell. The text below the photo reads: 'O presidente ficou 7 dias sem fazer m**** no governo e foi parar no hospital | Reprodução/ Redes sociais'. The article text continues: 'Num quadro que desafia a medicina, Jair Bolsonaro foi internado na noite de ontem após sentir um desconforto intestinal. Fontes dizem que ele sentiu enjoo depois de se lembrar que é o Jair Bolsonaro. "Olhou no espelho e começou a vomitar", disse um assessor.'

On the right side, there is a section titled 'OUTRAS PÁGINAS' with several links to other content: 'Patrícia Kogut' (Humorístico de Leandro Hassum irá ao ar também na Globo), 'Joaquim Ferreira dos Santos' (Nara é que era mulher de verdade), 'Blog do Bonequinho' (Ao mestre, com carinho), 'Blog do Acervo' (Raul Seixas de bicicleta: fotos inéditas do pai do rock brasileiro, em 1973), 'Afonso Borges' (Emediato, da editora Geração, é categorico: "Daniela Arbex mente em suas acusações"), 'Saideira' (Novidades: Cerveja lançada em 10 estados: 60 anos do Cadeg; holding Evinó e Grand Cru), 'Ruth de Aquino' (A guerra entre o Papa e as mães de pets), and 'Amplificador' ('Morena bonita': canção de Anna Setton ganha clipe que celebra amor entre duas mulheres).

Fonte: Página do Sensacionalista.⁷

4.1.2 Criação do SatiriCorpus.Br

O SatiriCorpus.Br é um *corpus* constituído por notícias satíricas do português do Brasil. Para a construção do SatiriCorpus.Br, um *crawler* coletou automaticamente as notícias do site Sensacionalista, coletando o texto do corpo da notícia que estava na página, excluindo ruídos como *tags*, *html*, figuras. Estes arquivos foram salvos em um arquivo JSON⁸, que tinha um campo “texto”, o que originou cada notícia no formato de texto. Vale ressaltar que o período de extração do *corpus* foi em janeiro de 2019 e, desse modo, considerou-se o início das postagens das notícias no site do Sensacionalista – em 2016 até o final de 2018. Obedecendo a própria classificação temática que o site estabelecia⁹, o *corpus* foi dividido em cinco categorias: i) *comportamento*, ii) *entretenimento*, iii) *esporte*, iv) *mundo* e v) *país*; descritas no Quadro 4. Já a Tabela 1 descreve os dados quantitativos de cada categoria quando ao total de notícias e de *tokens*.

⁷ Disponível em: <<https://blogs.oglobo.globo.com/sensacionalista/>>. Acesso em: 08 de jan. 2022.

⁸ JSON, um acrônimo para *JavaScript Object Notation*. É basicamente um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas.

⁹ Na época da extração, a hospedagem do site Sensacionalista pertencia a Veja. No ano de 2021, o site foi desativado e passou a ser uma coluna de humor no O Globo.

Quadro 4 – Descrição dos assuntos abordados nas categorias

Categorias	Descrição dos assuntos
Comportamento	Sobre comportamentos e atitudes humanas; cotidiano da sociedade.
Entretenimento	Sobre celebridades; assuntos da TV brasileira e internacional.
Esporte	Sobre esporte brasileiro e internacional.
Mundo	Sobre política internacional.
País	Sobre política brasileira.

Fonte: Elaborado pela autora.

Tabela 1 – Características do *corpus*

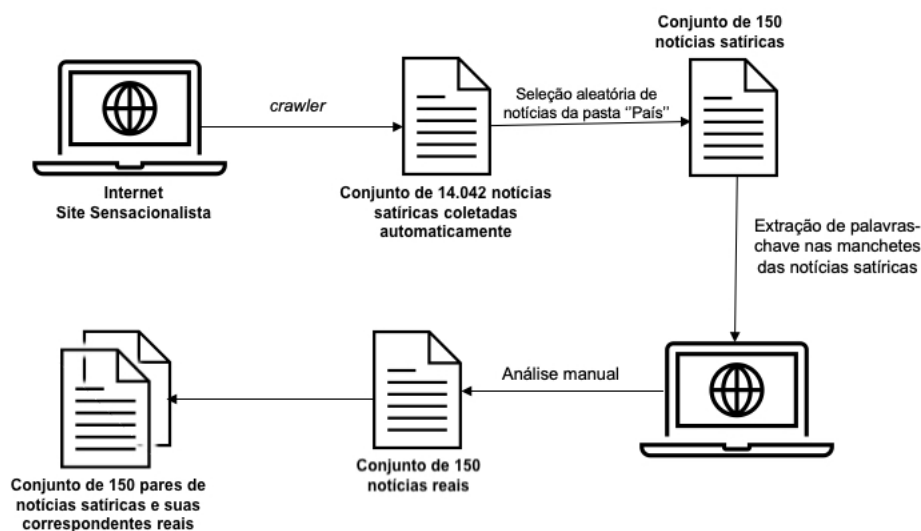
Categorias	Número de notícias	Tokens	Types	Sentenças	
Comportamento	56	1,11%	8.014	6.947	267
Entretenimento	1.406	27,86%	255.972	218.531	10.720
Esporte	738	14,61%	120.332	103.568	4.931
Mundo	1.001	19,83%	178.185	152.555	8.250
País	1.847	36,60%	316.588	271.293	12.347
Total	5.048		879.091	753.293	36.515

Fonte: Elaborada pela autora.

Uma vez que um dos principais objetivos da pesquisa é a investigação de elementos linguísticos e a descrição linguística de notícias satíricas, optou-se pela construção de um *subcorpus* formado por notícias satíricas e seu paralelo de notícias verdadeiras. O propósito é compreender como os fenômenos linguísticos ocorrem em textos contendo informações factuais em comparação aos textos satíricos.

4.1.3 Construção do *subcorpus*

Ao considerar os requisitos de um *corpus* de notícias falsas propostos por (RUBIN; CHEN; CONROY, 2015), quando os autores estabelecem que o alinhamento das notícias falsas com as verdadeiras é importante para verificar as instâncias positivas e negativas para validar padrões linguísticos, o *corpus* descrito na seção anterior foi segmentado em um *subcorpus* constituído por 150 notícias selecionadas arbitrariamente da categoria “País” e 150 notícias verdadeiras referentes às notícias satíricas. Em relação às notícias verdadeiras, primeiramente, delimitaram-se palavras-chave identificadas nas notícias satíricas. Em seguida, com estas palavras-chave estabelecidas, foi elaborada uma busca manual das notícias verdadeiras equivalente às satíricas. Para evitar a seleção de notícias falsas, o critério era selecionar apenas notícias veiculadas a tradicionais portais de notícias online. É importante destacar que a escolha pela categoria “País” se dá pelo fato de conter majoritariamente notícias sobre a política brasileira, o que ajudaria na identificação de palavras-chave e na busca pelas notícias verdadeiras. O processo é descrito na Figura 15.

Figura 15 – Processo de construção do *subcorpus* de análise

Fonte: Elaborada pela autora.

Os dados presentes nas características do *subcorpus* descritas na Tabela 2 foram gerados a partir do NLTK¹⁰, (do inglês, *Natural Language ToolKit*), uma biblioteca *open source*¹¹ de ferramentas úteis na linguagem Python¹² para a utilização dos princípios de PLN e pelo spaCy¹³, que também é uma biblioteca desenvolvida para Python para processamento de língua natural. Sua utilização é para uso em produção e para ajudar a criar aplicações que conseguem processar e “entender” um grande volume de texto.

Tabela 2 – Características do *subcorpus* (NLTK)

Descrição	Reais			Satíricas		
	Total	Média	Desvio Padrão	Total	Média	Desvio Padrão
Número de <i>tokens</i>	107.133	714,22	570,42	22.963	153,08	46,19
Número de <i>types</i>	11.304	299,5	170,49	14.843	98,95	26,75
Número de sentenças	5.721	38,14	30,27	1.246	8,08	2,51
Número de caracteres	651.568	4.343,78	3.519,51	135.966	906,44	277,92
Número de sílabas	231.195	1.541,30	1.252,26	48.234	321,56	97,97

Fonte: Elaborada pela autora.

Os números de *tokens* e de *types* foram extraídos do NLTK. O processo de extração é ilustrado pela Figura 16.

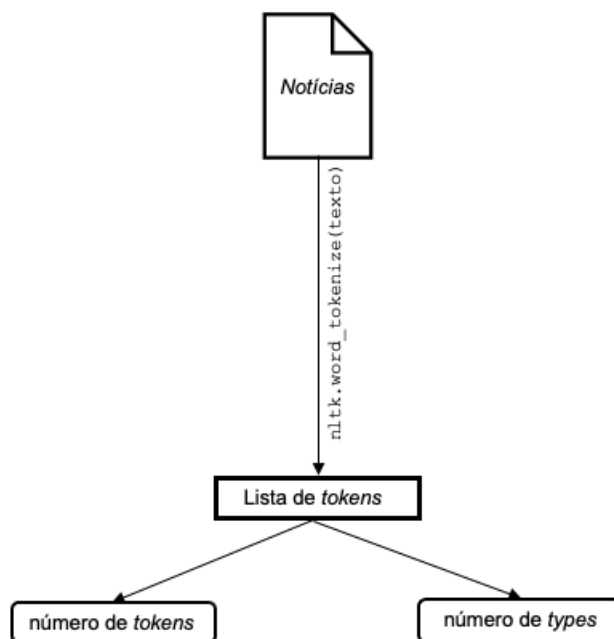
¹⁰ Disponível em: <<https://www.nltk.org/index.html>>.

¹¹ Código aberto: distribuição livre, código-fonte, trabalhos derivados, distribuição da licença, entre outros.

¹² Disponível em: <<https://www.python.org/>>.

¹³ Disponível em: <<https://spacy.io/>>

Figura 16 – Processo de extração de características pelo NLTK



Fonte: Elaboração da autora.

A tarefa de *tokenização* de um texto foi realizada a partir da linha de código `nltk.word_tokenize(texto)`. Essa função recebe o texto como argumento e retorna todas as palavras do texto (ou da sentença) em forma de *tokens*.

O *número de tokens* é a quantidade total de palavras do texto. Entende-se por *tokens* cada palavra, número ou sinal de pontuação presente no texto e *número de types* é referente à quantidade de palavras diferentes no texto. Assim, a sentença

(11) Até o pato da FIESP deve pagar o pato. [ex.s]

possui 10 *tokens* ('até', 'o', 'pato', 'da', 'FIESP', 'deve', 'pagar', 'o', 'pato') e 6 *types* ('até', 'o', 'pato', 'da', 'FIESP', 'deve', 'pagar'), dado que 'pato' e 'o' são contabilizados apenas uma vez.

Os dados relativos ao número de caracteres, de sílabas e de sentenças foram gerados pelo spaCy. O *número de caracteres* contabiliza o número total de caracteres do texto. O *número de sílabas*, o total da extração de sílabas de uma única palavra do *subcorpus* real e satírico. A contagem de sílabas é utilizada no cálculo do Índice Flesch (cf. na Seção 5.2.3).

O *número de sentenças* se refere à quantidade de sentenças de cada notícia do *subcorpus*. Entende-se por sentença o segmento do texto iniciado por letra maiúscula e terminado por ponto final, ponto de interrogação, ponto de exclamação ou reticências.

Os dados presentes na Tabela 3 foram extraídos do NILC-Metrix (LEAL, 2021)¹⁴ (cf. Seção 5.2.3).

¹⁴ Disponível em: <<http://fw.nilc.icmc.usp.br:23380/metrixdoc>>.

Tabela 3 – Características do *subcorpus* (NILC-Metrix)

Descrição	Notícias	
	Reais	Satíricas
Riqueza lexical (TTR)	0,73	0,73
Média de palavras por sentença	18,35	19,57
Média de verbos	19,22	23,51
Média de verbos modais	2,59	3,21
Média de substantivos	37,12	37,28
Média de adjetivos	5,90	5,67
Média de advérbios	5,40	7,43
Média de pronomes	0,36	0,39

Fonte: Elaborada pela autora.

A riqueza lexical (TTR)¹⁵ e a média de palavras, verbos, verbos modais, substantivos, adjetivos, advérbios e pronomes foram extraídos pelas métricas descritivas do NILC-Metrix. A *média de palavras por sentença* é estabelecida a partir do número médio de palavras por sentença no texto. Já o *TTR* está relacionado à diversidade lexical do texto, sendo calculado pela razão dos *types* e o número de *tokens*. O cálculo da *média dos verbos*, dos *verbos no imperativo* e da *média de verbos modais* é realizado entre a divisão da quantidade de verbos, verbos no imperativo e verbos modais, respectivamente, e o número de *tokens*. A *média de substantivos* é a relação entre o número total de substantivos em relação ao número total de palavras, assim como a *média de adjetivos* é a relação dos adjetivos ao número total de palavras do texto. Seguindo, a *média de advérbios* é a divisão entre os advérbios do texto e os tokens e a *média de pronomes* é calculada entre o número de pronomes pelo número de totais de palavras da notícia.

Nas médias de verbos, verbos modais, substantivos, adjetivos, advérbios e pronomes foi realizada uma normalização.¹⁶ pela quantidade de *tokens* para evitar ter uma média (de verbos, por exemplo) muito maior na coluna das notícias reais, porque tem mais *tokens*. Como aponta Finatto (2011, p. 9), a normalização das frequências tem o objetivo de nivelar a extensão irregular dos textos. Por exemplo, na média de verbos, existem em média 19,22 verbos quando analisadas todas as notícias reais e 23,51 verbos quando analisadas todas as notícias verdadeiras, normalizadas pela quantidade de *tokens*. Ou seja, aqui é possível comparar como se as notícias reais e satíricas possuíssem a mesma quantidade de palavras. É importante ressaltar que as médias foram calculadas automaticamente pelo NILC-Metrix em cada notícia, resultando uma média total das 150 notícias reais e satíricas.

Nota-se uma pequena diferença estatística entre as notícias, com exceção à média de verbos e média de advérbios, ambas mais frequentes nas notícias reais. Isso acontece

¹⁵ Do inglês, *Type-Token Ratio*.

¹⁶ Considera-se normalização como a média de uma categoria (verbos, substantivos, por exemplo) pela quantidade total de *tokens* do texto.

pelas notícias reais serem maiores textualmente do que as notícias satíricas e embora se compreenda a importância da homogeneidade do tamanho do texto, principalmente ao considerar abordagens de aprendizado de máquina – apontado por Rubin, Chen e Conroy (2015) – nesta pesquisa, decidiu-se por não balancear o *corpus* em relação ao tamanho de cada notícia satírica e real. Como esta tese é baseada em uma perspectiva linguística, entende-se que o não balanceamento do *corpus* pode ser útil para evitar a perda de informações na análise, uma vez que o número de palavras, sentenças ou diversidade lexical pode ser uma característica para a diferenciação desse tipo de conteúdo.

4.2 ASPECTOS LINGUÍSTICOS NA DESCRIÇÃO DAS NOTÍCIAS SATÍRICAS

A partir da construção do *subcorpus*, realizou-se uma análise linguística, procurando observar os principais recursos linguísticos na construção do efeito satírico nas notícias. Desse modo, esta seção se ocupa em apresentar as etapas realizadas para a análise do *subcorpus* satírico.

4.2.1 Classificação das sentenças do *subcorpus*

A primeira etapa realizada foi a divisão dos textos que compõem o *subcorpus* em sentenças definidas como a unidade mínima de análise segmentada por um ponto final, exclamação, interrogação ou reticências. Para facilitar a visualização, os dados foram alocados em planilhas distintas: uma para as sentenças satíricas e outra para as reais. Na planilha das notícias satíricas, as sentenças foram analisadas manualmente e separadas em quatro categorias: manchetes¹⁷, sentenças verdadeiras, sentenças satíricas e discurso direto. As manchetes são as primeiras sentenças das notícias e o discurso direto refere-se àquelas em que se reproduz as palavras de outras pessoas e não a do autor da notícia.

- **Manchete na notícia satírica:**

(12) Bolsonaro não vai a debates porque adolescente não assiste TV aberta. [ex.s]

- **Sentença verdadeira na notícia satírica:**

(13) O candidato a Presidência, Jair Bolsonaro, ainda não bateu o martelo se participará ou não de debates eleitorais e até de sabatinas. [ex.s]

- **Sentença satírica:**

(14) Outro motivo de Bolsonaro não querer participar de debates é seu medo de ter que falar mais de um minuto sobre economia. [ex.s]

¹⁷ As manchetes das notícias satíricas sempre serão satíricas.

- **Discurso direto na notícia satírica:**

- (15) “Adolescente não assiste TV aberta, tá ok? Então para que eu vou participar desses debates?”, disse Bolsonaro. [ex.s]

Para diferenciar as classificações das sentenças verdadeiras e satíricas, utilizou-se como base as notícias reais do *subcorpus* e também a Web. Essa categorização pode esclarecer quais dispositivos satíricos estão na estrutura de cada tipo dessas sentenças. Na planilha das notícias reais, as sentenças foram divididas em três categorias: manchetes, sentenças e discurso direto.

- **Manchetes na notícia real:**

- (16) Jair Bolsonaro afirma que não vai a debates no segundo turno. [ex.r]

- **Sentenças verdadeiras:**

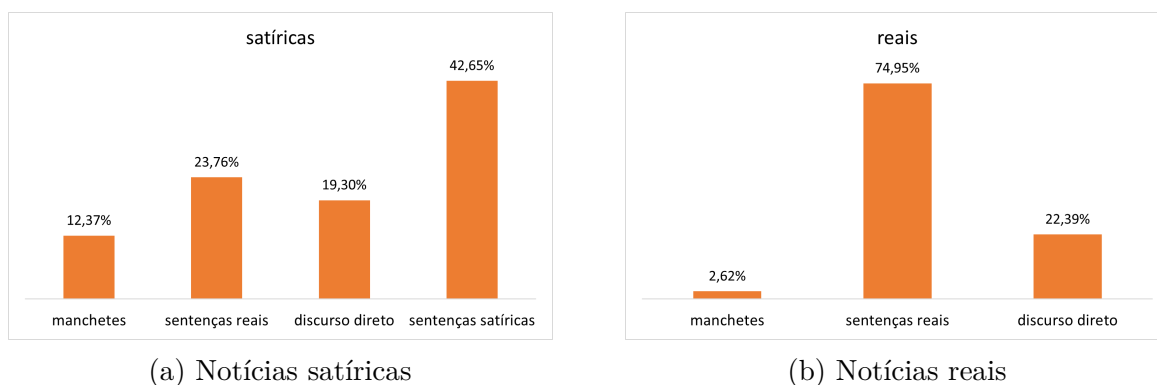
- (17) O candidato do PSL à Presidência, Jair Bolsonaro, disse no fim da tarde desta quinta-feira que não vai a debates marcados para o segundo turno das eleições. [ex.r]

- **Discurso direto na notícia real:**

- (18) “Segundo fui informado tenho restrições, eu poderia me submeter a uma aventura, de participar de um debate, de duas ou três horas, mas poderia ter uma consequência péssima para minha saúde.” [ex.r]

A Figura 17a apresenta a disposição das sentenças nas notícias satíricas, enquanto a Figura 17b é referente às notícias reais.

Figura 17 – Disposição de sentenças no *subcorpus*



Fonte: Elaborada pela autora.

É importante salientar que as sentenças foram separadas em manchetes, sentenças (satíricas e reais) e discurso direto por acreditar que a estrutura e construção da sátira

difira em cada tipo de sentença. Dessa maneira, as manchetes são mais curtas que as sentenças e o discurso direto, além de provavelmente estar na 1ª pessoa, o efeito satírico é muito mais sutil quando comparado às outras sentenças.

4.2.2 Classificação da interpretação satírica: explícita e implícita

Como já dito anteriormente, a compreensão da sátira, na maioria das vezes, está além dos elementos linguísticos presentes na superfície do texto, o que exige, portanto, um contexto extralinguístico e um conhecimento de mundo do leitor para que a sátira seja bem-sucedida. Nesse contexto, Wick-Pedro (2018), ao analisar comentários irônicos de notícias políticas do português do Brasil, observou que as opiniões irônicas poderiam conter uma oposição explícita ou implícita. Ou seja, como a ironia é frequentemente entendida por meio da oposição e contradição de elementos subjacentes no texto, essa oposição pode ser uma oposição entre elementos lexicais (explícitas) ou contextuais (implícitas). Aqui, como a ironia não é objeto principal do estudo (apesar de ser considerada fortemente como um elemento essencial da sátira) é feita uma adaptação do conceito das oposições explícitas e implícitas: **interpretação explícita** e **interpretação implícita**, como se observa no Quadro 5.

Quadro 5 – Sátira implícita e explícita

INTERPRETAÇÃO	
EXPLÍCITA	IMPLÍCITA
A sátira é compreendida por meio da relação de sinais presentes no enunciado.	A sátira é compreendida por meio de sinais que estão em um contexto pragmático adicional ao enunciado.

Fonte: Adaptado de Wick-Pedro (2018).

Assim, a interpretação explícita pode conter um contraste entre elementos, principalmente nos casos de contraposição de elementos que não se espera aparecer juntos por não pertencerem ao mesmo campo léxico-semântico, quebrando a expectativa do leitor, construindo o efeito de humor.

- (19) Temer vai ser contratado pelo **Sensacionalista** a partir de 2019. [ex.s]
 (20) **Alckmin** dá um **drible** na **Lava-Jato** e é **convocado por Tite**. [ex.s]

Nos exemplos acima, observa-se o uso de recursos linguísticos que podem indicar que a sentença é uma sentença satírica. Em 19, o uso de “*Sensacionalista*” é um indicativo de pertencer a uma notícia satírica. Isso ocorre por ser uma marca da autoria e da origem desse enunciado. Já no exemplo 20, os itens lexicais ‘*Alckmin*’, ‘*Lava-Jato*’ pertencem ao domínio semântico da política, enquanto ‘*Tite*’ e ‘*convocado*’ pertencem ao domínio semântico do futebol. O contraste fica evidente pela oposição destes domínios (político

versus futebol) com o uso da palavra ‘drible’ – palavra do domínio do futebol, mas que é aplicada, neste exemplo, com seu significado figurado: ação de desvencilhar-se com destreza de adversários.

No entanto, uma sentença satírica contendo uma interpretação implícita descreve quando a sentença é compreendida como satírica por meio de sinais em um contexto pragmático adicional ao enunciado. Assim, se a manchete satírica não contém elementos linguísticos presentes na estrutura do texto que indiquem o efeito satírico – seja um efeito cômico ou de humor, por exemplo, sendo necessário que o leitor possa detectar o conteúdo satírico com base no conhecimento de mundo em comum com o tema abordado na notícia satírica.

- (21) Áudio de Joesley já comprovava que Temer é muito vivo. [ex.s]
- (22) Gilmar Mendes já abre a geladeira sem pedir e usa banheiro de porta aberta no Jaburu. [ex.s]

Os cálculos para a compreensão da sátira nos exemplos (21) e (22) é muito mais complexo se comparado aos exemplos anteriores. Em (21), é necessário que o leitor reconheça primeiramente a comparação entre o ex-presidente Michel Temer com um vampiro¹⁸, para depois relacioná-lo ao áudio de Joesley Batista, que discutiam um esquema de corrupção. O exemplo 22 se refere ao evento sobre a aproximação do ministro Gilmar Mendes com o ex-presidente Michel Temer em 2017 e sobre as visitas do ministro ao ex-presidente no Palácio do Jaburu¹⁹.

Existem casos em que a sátira pode ser inferida a partir de mecanismos presentes na superfície textual, sinalizando que aquele enunciado pode ser satírico. Logo, quando o efeito satírico é explicitamente ativado, considera-se como uma interpretação explícita. Todavia, quando existe uma camada ‘oculta’ e mais profunda, que exige do leitor mais cálculos semânticos e conhecimento de mundo e contextual para alcançar o sentido satírico do enunciado, entende-se como uma interpretação implícita. No entanto, a distinção parece nem sempre ser clara sobre o conteúdo implícito ou explícito em um texto.

- (23) **Foro privilegiado** cai e renova a esperança do fim da **tomada de três pinos**. [ex.s]

Em (65) há uma contraposição entre ‘*tomada de três pinos*’, que não tem nenhuma relação semântica com ‘*foro privilegiado*’, pertencente ao domínio político ou jurídico, indicando a construção de um efeito humorístico no enunciado. Porém, esse efeito de humor só pode ser considerado quando leitor tem o conhecimento de mundo sobre a adoção da tomada de três pinos no Brasil.

¹⁸ Disponível em: <<https://bit.ly/3ErmvXW>>. Acesso em: 13 mai. 2022

¹⁹ Disponível em: <<https://bit.ly/3ECIET8>>. Acesso em: 13 mai. 2022

4.2.2.1 Anotação de sentenças explícitas e implícitas

De acordo com Hovy e Lavid (2010), a anotação de *corpus* é um processo em que os humanos – ou anotadores – podem adicionar novas informações em dados brutos ao texto. Estas informações são adicionais por decisões cognitivas que dependem tanto dos textos brutos quanto de alguma teoria ou conhecimento que o anotador já tenha internalizado anteriormente. Segundo os autores, é necessário, primeiramente, criar diretrizes que garantam que todos os anotadores realizem uniformemente toda a tarefa de anotação. Além das diretrizes, uma anotação deve também respeitar o seguinte esquema:

1. Definir a tarefa de anotação com base na necessidade da pesquisa.
2. Selecionar os dados a serem anotados.
3. Escrever um conjunto detalhado de diretrizes da anotação.
4. Criar e utilizar boas ferramentas de anotação.
5. Encontrar e treinar anotadores.
6. Anotar o texto:
 - a) Anotar o texto com base nas diretrizes.
 - b) Revisar as diretrizes de anotação, se necessário.
 - c) Monitorar a concordância inter-anotador e retreinar os anotadores.
 - d) Se necessário, modificar a anotação com base nas diretrizes revisadas.
7. Tornar disponível o *corpus* para outras comunidades de pesquisa.

Assim como Hovy e Lavid (2010) já haviam apontado, Pustejovsky e Stubbs (2012) afirmam que ao fornecer dados para anotação, é necessário avaliar sobre o quão confiável é a anotação. Para isso, os autores propõem uma medida para comparar a anotação dos anotadores individuais entre si: a concordância inter-anotador (do inglês, *inter-annotator agreement* ou IAA). Assim, o IAA pode ser aplicável para validar esquemas e diretrizes de anotação, identificar ambiguidades ou dificuldades na fonte ou avaliar o alcance de interpretações úteis (PUSTEJOVSKY; STUBBS, 2012; ARTSTEIN, 2017). A Figura 18 ilustra o processo proposto por Pustejovsky e Stubbs (2012).

Figura 18 – Modelo de anotação



Fonte: Adaptado de Pustejovsky e Stubbs (2012, p. 25)

O coeficiente mais utilizado é o *kappa* (κ). O coeficiente *kappa* de Cohen avalia o acordo de não mais que dois anotadores, enquanto o *kappa* de Fleiss é uma medida estatística que avalia o nível de concordância ou reprodutibilidade entre dois, ou mais anotadores ao atribuir classificações a um conjunto de dados (PUSTEJOVSKY; STUBBS, 2012). Por esta pesquisa considerar a anotação de três anotadores, a verificação da concordância inter-anotador da anotação será através do *kappa* de Fleiss, detalhada no próximo capítulo.

Entende-se que seguir os esquemas apresentados por Hovy e Lavid (2010) e Pustejovsky e Stubbs (2012) é, idealmente, o processo necessário para a produção de uma anotação de alta qualidade. Pustejovsky e Stubbs (2012, p. 25) ainda salientam que “tarefa ainda pode precisar ser revisada, mesmo que suas pontuações no IAA sejam altas”²⁰ No entanto, por ser uma anotação desenvolvida para validar um método analisado, não foram realizadas as etapas de treinamento de anotadores, revisão das diretrizes e nem o monitoramento do acordo inter-anotador.

Assim, como já dito anteriormente, a compreensão da sátira, na maioria das vezes, está além dos elementos linguísticos presentes na superfície do texto, o que exige, portanto, um contexto extralinguístico e um conhecimento de mundo do leitor para que a sátira seja bem-sucedida. No entanto, podem existir mecanismos linguísticos presentes na estrutura textual que sinalizem o conteúdo satírico do texto. Logo, em princípio, para validar a classificação de interpretação externa ou interna à língua, foram criadas as diretrizes que norteiam esta anotação. A diretriz completa consta no Apêndice A. A análise foi feita por três anotadores. Cabe ressaltar que todos os anotadores são linguistas, com conhecimento sobre política e sabem que se trata de um conteúdo satírico.

Para tanto, cada anotador recebeu um formulário *online* contendo as 150 manchetes das notícias satíricas, como mostra a Figura 19. As manchetes deveriam ser classificadas de acordo com categorias **explícitas** ou **implícitas**. Detalhadas como:

1. Explícita: A sátira é compreendida por meio da relação de sinais existentes no

²⁰ “[...] task may still need to be revised even if your IAA scores are high.”

enunciado. Assim, se a manchete satírica contém elementos linguísticos presentes na estrutura do texto que indiquem o efeito satírico – seja um efeito cômico ou de humor, por exemplo – ela deverá ser interpretada como **sentença satírica explícita**.

2. Implícita: A sátira é compreendida por meio de sinais em um contexto pragmático adicional ao enunciado. Assim, se a manchete satírica **NÃO** contém elementos linguísticos presentes na estrutura do texto que indiquem o efeito satírico – seja um efeito cômico ou de humor, por exemplo, é necessário que o leitor que compreenda o conteúdo satírico com base no conhecimento de mundo em comum com o tema abordado na notícia satírica. Quando isso ocorrer, ela deverá ser interpretada como **sentença satírica implícita**.

Figura 19 – Exemplo do formulário de anotação

	A	B	C	D	E	F
1	Manchetes	Explícita	Implícita	Comentários		
2	Bolsonaro não vai a debates porque adolescente não assiste TV aberta	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
3	PSDB homenageou Gilmar Mendes ontem porque ele é uma mãe para eles	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
4	Bom dia para Lula em Curitiba é mais chato que bom dia em grupo de família do Whatsapp, diz pesquisa	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
5	Lula perdoa PMDB e estamos cansados demais para fazer piada	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
6	Após anos de comerciais do Dollynho, dono da Dolly finalmente é preso	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
7	Suíça emite alerta e pode quebrar se tiver que repatriar dinheiro do PSDB	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
8	Joaquim Barbosa desiste da presidência e mira o comando do Caldeirão do Huck	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
9	Alerj vota para 17 de novembro ser o dia de 'santa Carmen Lúcia'	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
10	Após começar desempregando Dilma, Temer atinge a marca de 13,7 milhões sem emprego	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
11	Cachorro de Marcela se jogou no lago por ter que conviver com Temer	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
12	Foro privilegiado cai e renova a esperança do fim da tomada de três pinos	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
13	Temer fecha 500 agências dos Correios para evitar que Lula receba cartas	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
14	Temer vai ser contratado pelo Sensacionalista a partir de 2019	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
15	Com medo de ser preso, morador da Zona Oeste do Rio começa a levar advogado como par no pagode	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
16	PF pede transferência de Lula e 31% dos brasileiros sugerem Palácio do Planalto	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
17	Repórter que fazia 'povo fala' na rua pergunta a Bruno Covas o que ele acha do novo prefeito	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
18	Petista vence BBB e Moro já manda prender	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
19	Joaquim Barbosa estuda abafar algum escândalo para atrair eleitor do PSDB	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
20	Aécio se junta aos 51 milhões que votaram nele e diz que foi ingênuo	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
21	Gleisi Hoffmann investigada por ouvir disco da Ai Cione	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
22	Cirque du Soleil abre seleção no Brasil após exposições de malabarismo em defesa de Waack	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
23	Darth Vader processa Marina por dizer que Bolsonaro é do lado negro da Força	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
24	Após deixar triplex, MTST quer ocupar sítio em Atibaia mas ninguém se voluntariou	<input checked="" type="checkbox"/>	<input type="checkbox"/>			

Fonte: Elaborada pela autora.

4.3 FERRAMENTAS LINGUÍSTICO-COMPUTACIONAIS

Para o desenvolvimento da presente pesquisa foram utilizados alguns recursos computacionais, como o etiquetador morfossintático PALAVRAS (BICK, 2000), as métricas de inteligibilidade textual do NILC-Matrix (LEAL, 2021) e o dicionário e léxico do LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001; BALAGE FILHO; PARDO; ALUÍSIO, 2013) para auxiliar a extração de dados linguísticos para facilitar a análise e a identificação de padrões ou mecanismos linguísticos na construção da sátira nas notícias.

4.3.1 Etiquetação morfossintática pelo PALAVRAS

A etiquetação de *corpus* pode ser definida como o processo de enriquecer um *corpus*, adicionando informações linguísticas inseridas por humanos ou máquinas com um objetivo (teórico ou prático). Embora um *corpus* represente um recurso muito útil para estudos linguísticos, um *corpus* anotado significa para linguística um recurso linguístico ainda mais importante e valioso. Isso acontece porque as anotações acrescentam valor ao *corpus*, permitindo que buscas e processamentos mais refinados sejam realizados. Uma etiquetação

pode ser manual, quando realizada por linguistas; automática, quando processada por ferramentas de PLN e semiautomática, quando a correção da saída de outras ferramentas é manual.

Identificar classes morfossintáticas, classes gramaticais ou *parts-of-speech* (POS), termo mais utilizado no PLN, pode ser útil por descrever o comportamento de uma palavra ou das palavras vizinhas em um texto. Além disso, tais informações permitem a identificação de diversas questões relacionadas à estrutura sintática em que uma determinada palavra se encontra.

A atribuição automática de etiquetas morfossintáticas (em inglês, *POS tagging*), tem recebido muita atenção, pois é um componente fundamental para a maioria dos sistemas de PLN, uma vez que muitas tarefas exigem a atribuição de etiquetas que especifiquem as classes morfossintáticas das palavras (PETROV; DAS; MCDONALD, 2012; VOUTILAINEN, 2012). Os *POS taggers*, por sua vez, são as ferramentas computacionais que processam o texto e atribuem as etiquetas morfossintáticas a cada palavra.

Ainda, enquanto os *taggers* lidam com a estrutura das palavras e com suas classificações das classes gramaticais, o analisador sintático automático (em inglês, *parser*) trabalha com o conjunto de palavras e sua disposição nas sentenças. Desse modo, o *parsing* sintático é uma tarefa que consiste no reconhecimento automático da estrutura de uma sentença a partir de uma gramática para atribuir uma análise sintática, geralmente detalhada, a uma sequência de palavras com a finalidade de gerar árvores sintáticas (isto é, a saída de um *parser*, representada pela estrutura sintática em formato de árvore) que representem a estrutura sintática da oração analisada (CARROLL; MINNEN; BRISCOE, 2003; JURAFSKY; MARTIN, 2008).

O *subcorpus* foi submetido ao *parser* PALAVRAS²¹ (BICK, 2000), que gerou uma etiqueta para cada palavra, marcações sobre sua classe morfológica e seu papel sintático (cf. Seção 5.2.1. O PALAVRAS é um *parser* desenvolvido especialmente para o português). De maneira geral, o processo de anotação do *corpus* acontece nas duas seguintes etapas:

1. Primeiramente, o *corpus* – sem nenhuma anotação – é enviado para o analisador sintático PALAVRAS,
2. Em seguida, o PALAVRAS devolve o *corpus* anotado em relação às classes morfossintáticas e as funções sintáticas que cada palavra ocupa na sentença.

O conjunto de etiquetas do PALAVRAS contém 14 categorias de classes gramaticais (descritas no Quadro 6), que se combinam ainda com mais 24 etiquetas que descrevem a flexão de cada palavra, como:

- i) *Gênero*: masculino (M) ou feminino (F);

²¹ O *parser* PALAVRAS (BICK, 2000) não é de uso livre e exige aquisição da licença.

- ii) *Número*: singular (S) ou plural (P);
- iii) *Caso*: nominativo (NOM), acusativo (ACC), dativo (DAT) ou prepositivo (PIV);
- iv) *Pessoa*: primeira (1), segunda (2), terceira (3), da qual sempre está unida ao número, por exemplo, terceira pessoa do singular (3S);
- v) *Tempo verbal*: presente (PR), imperfeito (IMPF), perfeito simples (PS), mais-que-perfeito (MQP), futuro (FUT) e condicional (COND);
- vi) *Modo verbal*: indicativo (IND), subjuntivo (SUBJ) e imperativo (IMP) e
- vii) *Conjugação*: verbo finito (VFIN), infinitivo (INF), particípio (PCP) e gerúndio (GER).

Quadro 6 – Categoria das classes gramaticais do PALAVRAS

Etiqueta	Classe Gramatical	Exemplo
N	Nome	jornal, computador, bola
PROP	Nome próprio	Lula, Dilma, PT
SPEC	Especificadores	que, nada, quem
DET	Determinante	o, a, este, essa
PERS	Pronome	eu, lhe, mim
ADJ	Adjetivo	corrupto, esperta, constante
ADV	Advérbio	até, hoje, não, especificamente
V	Verbo	desistir, mirar, roubar
NUM	Numeral	oito, três, 13
PRP	Preposição	de, para, desde
KS	Conjunções subordinativas	se, que, porque
KC	Conjunções coordenativas	e, ou, mas
IN	Interjeições	oh!, ah!, uhul!
EC	Prefixos	anti-, vice-, ex-

Fonte: Elaborado pela autora.

A Figura 20 mostra a saída da anotação pelo *parser* PALAVRAS em duas sentenças retiradas do *subcorpus*: (a) ‘A pesquisa foi realizada por um instituto ligado ao PT’ e (b) ‘Havia boatos de que Paulo Preto assinaria acordo de delação premiada’, respectivamente.

Figura 20 – Exemplos da saída de uma sentença anotada pelo PALAVRAS

(a)

```

A [o] <*> <artd> DET F S @>N
pesquisa [pesquisa] <occ> <act-d> N F S @SUBJ>
foi [ser] <fmc> <aux> V PS 3S IND VFIN @FS-STA
realizada [realizar] <vH> <mv> V PCP F S @ICL-AUX
por [por] <PASS> PRP @<PASS
um [um] <arti> DET M S @>N
instituto [instituto] <inst> N M S @P<
ligado [ligar] <mv> <np-close> V PCP M S @ICL-N<
ao [ao] <sam-> <PIV> PRP @<PIV
PT [PT] <party> <*> PROP M S @P<

```

(b)

```

Havia [haver] <*> <fmc> <mv> V IMPF 1/3S IND VFIN @FS-STA
boatos [boato] <act-s> <ACC> N M P @<ACC
de [de] <np-close> PRP @N<
que [que] <clb> <clb-fs> KS @SUB
Paulo Preto [Paulo=Preto] <hum> <*> PROP M S @SUBJ>
assinaria [assinar] <vH> <mv> V COND 3S VFIN @FS-P<
acordo [acordo] <sem-c> <ACC> N M S @<ACC
de [de] <np-close> PRP @N<
delação [delação] <act> N F S @P<
premiada [premiado] <jh> <np-close> ADJ F S @N<

```

Fonte: Elaborada pela autora.

O processamento apresentado na imagem acima representa como todos os arquivos em texto simples foram gerados pelo *parser*. Na análise do verbo *realizada* da sentença (a), por exemplo, a sua saída começa com sua forma lematizada entre os colchetes (realizar) seguida de etiquetas secundárias, etiquetas semânticas com colchetes angulares para a maioria dos substantivos e verbos e alguns adjetivos. No exemplo, <vH> e <mv> representam verbo com sujeito humano e verbo principal, nessa ordem. Em azul, a etiqueta morfossintática indica o verbo no particípio (V, PCP), no feminino (F) do singular (S). Em verde, verifica-se a análise sintática que começa com o símbolo ‘@’. Vale ressaltar que quando um verbo pode ser analisado tanto na primeira pessoa do singular (1S) quanto na terceira pessoa do singular (3S), ele pode ser classificado em 1/3S, como no verbo haver na sentença (b).

4.3.2 Medição da inteligibilidade textual pelo NILC-Metrix

O dinamismo das mídias sociais possibilita uma maior rapidez de leitura, o que resulta em baixo aprofundamento de grande parte dos textos veiculados na rede e desqualifica o leitor para uma compreensão de textos mais complexos. Como aponta Idoeta (2019) em uma reportagem para a BBC Brasil, o fato de o leitor substituir o papel pelas telas propiciou o hábito de apenas “passar os olhos” superficialmente em diversos textos, interferindo na capacidade de compreender argumentos mais complexos ou então de realizar uma análise crítica do que está sendo lido. Para realizar o processo de leitura, o leitor precisa compreender totalmente o texto e ter o conhecimento prévio para obter integralmente o seu sentido (LEFFA, 1996). Esse cenário colabora para minimizar a capacidade de entender argumentos mais difíceis da linguagem, de fazer uma análise crítica do que está sendo lido, além de favorecer na ascensão da disseminação de notícias falsas.

Considera-se o texto como um resultado parcial da comunicação do leitor com processos cognitivos, contextuais e linguísticos (KOCH, 2013). De acordo com Resende e Souza (2011), o termo leiturabilidade (*readability*) refere-se ao que está inserido no ato de ler, considerando o papel do leitor, a habilidade, as características, os conhecimentos e a experiência do leitor na atividade de leitura. Já legibilidade (*legibility*) corresponde

aos “elementos e recursos que o próprio texto, em sua materialidade, oferece ao leitor” (RESENDE; SOUZA, 2011, p.3), i.e., relaciona-se com a construção do texto, das suas competências textuais que dão sustentação e materialização ao texto. Os aspectos físicos do texto, como a fonte escolhida, o espaçamento, as margens, etc. também se enquadram na legibilidade. Assim, a leiturabilidade de um texto tem a finalidade de calcular o nível de facilidade de leitura do leitor. Nesse sentido, entende-se que o tamanho das sentenças e o vocabulário do leitor aumentam (ou diminuem) a capacidade de leitura de um texto (DUBAY, 2004). Finatto e Paraguassu (2022, p. 43) apontam que apesar de Dubay (2004) não distinguir os termos leiturabilidade e inteligibilidade, no âmbito da computação, adota-se o termo inteligibilidade para se referir aos “textos que sejam mais simples de ler do que outros, ou seja, mais inteligíveis”. Logo, adota-se a mesma designação nesta pesquisa.

Nesta seção, abordam-se as medidas de inteligibilidade textual Coh-Metrix (GRAESSER et al., 2004; MCNAMARA et al., 2010), Coh-Metrix-Port (SCARTON; ALUISIO, 2010), Coh-Metrix-Dementia (ALUÍSIO; CUNHA; SCARTON, 2016), o NILC-Metrix (LEAL, 2021), que será aplicado aos textos de análise desta pesquisa, assim como o Índice Flesh para o português (MARTINS et al., 1996), uma adaptação do Índice Flesch de Facilidade de Leitura (FLESCH, 1981).

4.3.2.1 Índice Flesch

O Índice Flesch de Facilidade de Leitura (FLESCH, 1981) é uma fórmula que avalia, de modo superficial, a leiturabilidade de um texto em inglês. O Índice Flesch procura uma correlação entre tamanho da sentença e o tamanho da palavra a partir da fórmula:

$$206,835 - (1,015 \times \text{TMS}) - (84,6 \times \text{MSP})$$

em que TMS = tamanho médio das sentenças (o número de palavras dividido pelo número de sentenças) e MSP = média de sílabas por palavras (o número de sílabas divididas pelo número de sentenças).

O resultado da fórmula é um número entre 0 e 100, sendo 100-90 (muito fácil), 90-80 (fácil), 80-70 (muito fácil), 70-60 (padrão ou *plain language*²²), 60-50 (razoavelmente difícil), 50-30 (difícil) e 30-0 (muito difícil).

A fórmula de Flesch foi adaptada para o português por pesquisadores do Instituto de Ciências Matemática e da Computação da Universidade de São Paulo (ICMC-USP) (MARTINS et al., 1996). A fórmula corresponde ao Índice Flesch de Facilidade de Leitura somada com o número 42, pois de acordo com (MARTINS et al., 1996) é, na média, o número que diferencia os textos do inglês para textos em português. Portanto, a fórmula adaptada para o português é mostrada a seguir:

²² Linguagem simples.

248,835 - (1,015 x TMS) - (84,6 x MSP)

Os valores desse índice variam entre 100-75 (muito fácil), 75-50 (fácil), 50-25 (difícil) e 25-0 (muito difícil), correspondendo, respectivamente, do 1º ao 5º ano, 6º ao 9º ano, Ensino Médio e Ensino Superior.

4.3.2.2 *Coh-Metrix*

O Coh-Metrix (GRAESSER et al., 2004; MCNAMARA et al., 2010) significa *Cohesion Metrix* é uma ferramenta computacional desenvolvida para análise textual da língua inglesa, sendo um dos principais recursos utilizados para extrair métricas de inteligibilidade textual. O Coh-Metrix tem como propósito extrair métricas de coesão e coerência textual em medidas lexicais, sintáticas, semânticas e referenciais que avaliam a complexidade do texto.

O Coh-Metrix versão 3.0 implementa 106 métricas distribuídas em 11 categorias: medidas descritivas, pontuações dos principais componentes de facilidade de texto, coesão referencial, análise semântica latente (LSA), diversidade lexical, conectivos, modelo de situação, complexidade sintática, densidade de padrão sintático, informações de palavras e legibilidade.

4.3.2.3 *Coh-Metrix para o português*

A primeira ferramenta para avaliação de textos para o português do Brasil foi o Coh-Metrix-Port (SCARTON; ALUISIO, 2010), uma adaptação do inglês para a língua portuguesa. Desenvolvida no espaço do projeto PorSimples²³ (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), o Coh-Metrix-Port teve como principal objetivo promover o acesso a textos por crianças e/ou adultos analfabetos funcionais e/ou em fase de alfabetização. Esta versão trabalha com 48 métricas específicas para o português divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, *tokens*, constituintes conectivos, ambiguidade, correferência, anáforas, além de incluir o Índice Flesch.

Seguindo, o Coh-Metrix-Dementia²⁴ (ALUÍSIO; CUNHA; SCARTON, 2016) é uma adaptação do Coh-Metrix-Port, com foco na análise automática de distúrbios de linguagem relacionados à doença de Alzheimer e ao Comprometimento Cognitivo Leve. A ferramenta pode extrair 73 métricas, incluindo a adaptação das 48 métricas do Coh-Metrix-Port e mais 25 métricas referentes a disfluência, análise semântica latente, diversidade lexical, complexidade sintática e densidade semântica do texto.

²³ Disponível em: <<http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources?layout=edit&id=27>>

²⁴ Disponível em: <<http://143.107.183.175:22380/>>

Aprimorando o Coh-Metrix-Port (SCARTON; ALUISIO, 2010) e o Coh-Metrix-Dementia (ALUÍSIO; CUNHA; SCARTON, 2016), o NILC-Metrix²⁵ (LEAL, 2021) é um sistema computacional que contém por volta de 200 métricas propostas em estudos de discurso, psicolinguística, linguística cognitiva e computacional, com o objetivo de analisar a complexidade textual para o português.

Em geral, foram investigadas 200 medidas do NILC-Metrix, agrupadas em 14 categorias: *medidas descritivas, simplicidade textual, coesão referencial, coesão semântica, medidas psicolinguísticas, diversidade textual, conectivos, léxico temporal, complexidade sintática, densidade de padrões sintáticos, informações morfossintáticas de palavras, informações semânticas de palavras, frequência de palavras e índices de legibilidade*. Tais categorias vão desde informações morfossintáticas e frequência de palavras até medidas mais robustas, como medidas psicolinguísticas e de legibilidade e facilidade de leitura do texto. A documentação do sistema, bem como as explicações de todas as métricas estão disponíveis online²⁶.

4.3.3 LIWC

O LIWC, do inglês *Linguistic Inquiry and Word Count* (PENNEBAKER; FRANCIS; BOOTH, 2001), é um programa de análise computacional de texto que mede a quantidade de palavras de um texto, além de conter dicionários em vários idiomas, como inglês (PENNEBAKER et al., 2007; PENNEBAKER; FRANCIS; BOOTH, 2001), português (BALAGE FILHO; PARDO; ALUÍSIO, 2013), espanhol (SALAS-ZÁRATE et al., 2014), chinês (HUANG et al., 2012), entre outros. O LIWC foi desenvolvido especificamente para fornecer um método eficiente para estudar componentes emocionais, cognitivos e estruturais a partir de materiais textuais e tem sido utilizado para a classificação de dimensões psicológicas e estilo linguístico de conteúdo de mídias sociais (SALAS-ZÁRATE et al., 2014).

Considerando sua organização composicional, o LIWC pode ser dividido em duas partes – a primeira é o programa e a segunda são os dicionários. O programa, também conhecido como componente de processamento, abre uma série de arquivos de texto (nos mais diversos gêneros, como poema, romances, ensaios, blogs, etc.) e passa por cada um desses arquivos, analisando palavra por palavra. Assim, cada palavra analisada em um determinado arquivo de texto é comparada com o arquivo de dicionário que contém a palavra correspondente (TAUSCZIK; PENNEBAKER, 2009).

No dicionário, uma coleção de palavras é organizada em grupos de um domínio específico (PENNEBAKER et al., 2015). O dicionário contém palavras referentes a uma ou mais categorias que se referem a processos linguísticos, psicológicos ou relacionados a outros assuntos. De acordo com Tausczik e Pennebaker (2009), os dicionários são centrais

²⁵ Disponível em: <<http://fw.nilc.icmc.usp.br:23380/nilcmetrix>>

²⁶ Disponível em: <<http://fw.nilc.icmc.usp.br:23380/metrixdoc>>

para o LIWC. Em sua última versão, o LIWC conta com quase 6.400 palavras completas, seus *stems* e alguns *emoticons*. As palavras estão divididas em quatro grandes categorias: variáveis de sumarização da língua, dimensões linguísticas, processos psicológicos e outros tipos gramaticais. Ainda, cada uma destas categorias possui outras categorias, as quais contém um total de 92 subcategorias. Assim, ao enviar um texto para o LIWC, ele pode retornar o resultado que, por exemplo, 8,51% de um texto se enquadra na categoria de certeza do LIWC, enquanto apenas 0,81% se enquadra na categoria de religião e assim por diante.

A Figura 21 mostra como esta hierarquia de categoria é disposta no LIWC, especificamente, como a categoria processos psicológicos (*psychological process*), uma supercategoria, processos sociais (*social processes*) é uma categoria que contém 455 palavras relacionadas nesta categoria, e família (*family*) é uma subcategoria que contém 64 palavras relacionadas nesta subcategoria.

Figura 21 – Exemplo de hierarquia *psychological process* do LIWC

Psychological Processes					
Social processes ^b	social	Mate, talk, they, child	455		.97/.59
Family	family	Daughter, husband, aunt	64	.87	.81/.65
Friends	friend	Buddy, friend, neighbor	37	.70	.53/.12
Humans	human	Adult, baby, boy	61		.86/.26
Affective processes	affect	Happy, cried, abandon	915		.97/.36
Positive emotion	posemo	Love, nice, sweet	406	.41	.97/.40
Negative emotion	negemo	Hurt, ugly, nasty	499	.31	.97/.61
Anxiety	anx	Worried, fearful, nervous	91	.38	.89/.33
Anger	anger	Hate, kill, annoyed	184	.22	.92/.55
Sadness	sad	Crying, grief, sad	101	.07	.91/.45
Cognitive processes	cogmech	cause, know, ought	730		.97/.37
Insight	insight	think, know, consider	195		.94/.51
Causation	cause	because, effect, hence	108	.44	.88/.26
Discrepancy	discrep	should, would, could	76	.21	.80/.28
Tentative	tentat	maybe, perhaps, guess	155		.87/.13
Certainty	certain	always, never	83		.85/.29
Inhibition	inhib	block, constrain, stop	111		.91/.20
Inclusive	incl	And, with, include	18		.66/.32

Fonte: (PENNEBAKER et al., 2007, p.5).

De acordo com (TAUSCZIK; PENNEBAKER, 2009), o LIWC é um programa projetado para medir o significado social e psicológico das palavras. Pelas suas características psicolinguísticas, a principal motivação para o uso do LIWC é sua aplicação bem sucedida no que diz respeito à linguagem figurada (SALAS-ZÁRATE et al., 2014; SALAS-ZÁRATE et al., 2017). Para os autores, pela ferramenta conter um amplo conjunto de recursos, ela permite identificar diferenças entre a linguagem figurada e a linguagem literal. Nesta tese, usa-se a versão para o português (BALAGE FILHO; PARDO; ALUÍSIO, 2013), com 64 categorias e 127.160 palavras, seguindo a mesma estrutura do LIWC para o inglês. Para realizar a análise dos materiais extraídos do LIWC, foram gerados arquivos de texto simples com a análise da categoria processo psicológico (*psychological process*).

Desse modo, procurando observar as relações psicolinguísticas presentes nos textos tanto das notícias satíricas, quanto das notícias verdadeiras, as categorias empregadas do processo psicológico nesta pesquisa são as seguintes:

- **família** (*family*), como vitória, sucesso, melhor;
- **afeto** (*affect*), como feliz, chorou;
- **raiva** (*anger*), como odiar, matar, irritado;
- **ansiedade** (*anx*), preocupado, com medo;
- **certeza** (*certain*), sempre, nunca;
- **morte** (*death*), como enterrar, caixão, matar;
- **discordância** (*discrep*), deveria, seria;
- **sentir** (*feel*), sente, toque;
- **amizade** (*friend*), como amigo, vizinho;
- **ouvir** (*hear*), como ouvir, escutar;
- **inibição** (*inhib*), como bloquear, restringir;
- **intuição** (*insight*), pense, saiba;
- **lazer** (*leisure*), como cozinhar, conversar, filme;
- **religião** (*relig*), como altar, igreja;
- **tristeza** (*sad*), como chorando, pesar, triste;
- **palavrões** (*swear*), como foda, porra, merda;
- **tentativa** (*tent*), como pode ser, talvez, acho;
- **polaridade positiva**, como feliz, bonito, bom e
- **polaridade negativa**, como ódio, inútil, inimigo.

A escolha destas categorias ocorreu por se tratar de assuntos do cotidiano da sociedade e com forte relação com aspectos emocionais e afetivos, podendo aparecer de modos diferentes entre o texto real e o satírico.



Neste capítulo foram detalhadas as etapas da construção do SatiriCorpus.Br, um *corpus* de notícias satíricas para o português do Brasil e do *subcorpus*, uma fração de 150 notícias satíricas do *corpus* com 150 notícias reais paralelas. Discorreu-se também sobre as ferramentas de PLN que foram usadas para auxiliar a análise e busca de indícios de elementos linguísticos presentes nas notícias satíricas. O próximo capítulo apresenta uma análise linguística dos fenômenos linguísticos, tendo como base o *subcorpus* apresentado neste capítulo.

5 ANÁLISE LINGUÍSTICO-COMPUTACIONAL DE NOTÍCIAS SATÍRICAS

Para compreender melhor os fenômenos linguísticos que ocorrem nas notícias satíricas, foi preciso analisar o *subcorpus* satírico e descrevê-lo a partir de análises linguísticas. Essa descrição é relevante não só para o desenvolvimento desta tese, mas para uma melhor compreensão da construção da sátira e de outros recursos da linguagem presentes nesses textos. A partir da identificação desses fenômenos presentes nas notícias satíricas, é possível criar um mapeamento das ocorrências, criando, assim, uma tipologia de características linguísticas desses recursos.

Dessa forma, o capítulo apresenta dois tipos de análise linguística: i) manual, quando são identificadas características linguísticas no texto a partir de uma análise humana e uma leitura acurada das notícias satíricas e ii) computacional, quando são encontradas outras características linguísticas presentes no texto satírico por meio de ferramentas linguístico-computacionais. Assim, na Seção 5.1 é apresentada uma análise linguística e manual dos fenômenos que ocorrem no *subcorpus* das notícias satíricas e na Seção 5.2, apresenta-se a análise dos resultados obtidos pelo NILC-Matrix, que mediu a inteligibilidade textual das notícias verdadeiras e satíricas e pelo LIWC, que mediu a ocorrência de diversas categorias de palavras nos textos verdadeiros e satíricos.

Finalmente, na Seção 5.3 é analisado o processo de anotação das sentenças satíricas, anotadas em implícitas e explícitas. Também é descrito o processo de cálculo da concordância entre os anotadores.

5.1 ANÁLISE HUMANA NA DESCRIÇÃO DAS NOTÍCIAS SATÍRICAS

Esta seção descreve as características linguísticas encontradas a partir de uma análise manual e humana. Vale ressaltar que para compreender o funcionamento e construção das notícias satíricas, acredita-se ser necessário entender como o humano identifica os recursos linguísticos presentes na estrutura do texto satírico. Assim, a pesquisa não teve em vista encontrar uma determinada característica pré-estabelecida, mas estabelecer características a partir da análise. Este processo de análise linguística procurou não apenas buscar características, mas observar o efeito satírico e, então, classificá-lo em uma característica linguística identificada.

5.1.1 Características lexicais

Assim como apontaram Burfoot e Baldwin (2009) e Carvalho et al. (2020), a escolha lexical mais informal pode ser um indício de que a notícia é satírica, uma vez que os meios de comunicação dificilmente se utilizam de uma linguagem coloquial. Em (24), o item lexical *bolsominions* não ocorre nas notícias reais – embora seja frequente nas mídias sociais – pois, trata-se de uma **palavra pejorativa** para qualificar aqueles que estão

alinhados às ideias de Jair Bolsonaro. No exemplo (25), *pescotapa* pode ser entendido como um termo informal definido como uma “pancada dada com a mão na parte de trás do pescoço”¹

(24) **Bolsominions** já estão tomando toddynho com leite de ornitorrinco amazônico. [ex.s]

(25) Ciro caiu na provocação do blogueiro do MBL Arthur do Val e deu um **pescotapa** no rapaz. [ex.s]

O emprego de *foda* no exemplo (26) sugere a possibilidade de o texto não pertencer a uma notícia verdadeira, visto que jornais de grande circulação tendem a evitar **vulgarismos** (palavras grosseiras e ofensivas) em sua redação.

(26) Paulo é tão **foda** que Despacito é que fica com Paulo na cabeça. [ex.s]

Em uma busca pelos textos reais, visando encontrar a ocorrência de palavrões, encontrou-se a palavra *merda* tanto no *subcorpus* de notícias reais quanto no de notícias satíricas. Entretanto, é possível observar que em (27), o termo aparece dentro de um discurso direto, não pertencendo, portanto, ao texto do jornalista. O mesmo não acontece nas notícias satíricas, como demonstra o exemplo (28).

(27) “Tu acha que, se eu tivesse batido, não tinha uma marquinha, não? Do jeito que eu sou? Eu falei deixa de ser um **merda**, rapaz, e saí de perto”, disse. [ex.r]

(28) A última pesquisa eleitoral mostrou que Lula tem 35%, Bolsonaro tem 15% e o brasileiro tem **merda** na cabeça. [ex.s]

A marca de **autoria** satírica, presente nos exemplos (29) e (30), é o aspecto mais concreto, pois subscrevendo a fonte da notícia, dificilmente ela será reconhecida como uma notícia verdadeira.

(29) Para causar mais surpresa aos leitores, o **Sensacionalista** avaliou a possibilidade de passar a publicar notícias reais e sérias sobre a política brasileira. [ex.s]

(30) **Sensacionalista** pode encerrar atividades após Temer não provar que está vivo. [ex.s]

Já o exemplo (31) mostra uma notícia verdadeira sobre a confusão entre a *Al-Jazeera* e a *Al Qaeda* por grupos da direita. Assim, no exemplo (32) o humor está no trocadilho sonoro de *Al Caparras* e *Al Capone*. O **jogo de palavras** pode ser um forte indicativo de um enunciado satírico.

(31) O vídeo foi amplamente divulgado, principalmente por páginas de direita, que confundiram a rede de televisão **Al-Jazeera** com o grupo terrorista **Al Qaeda**. [ex.r]

¹ Disponível em: <<https://dicionario.priberam.org/pescotapa>>. Acessado em: 13.out.2021.

- (32) Ela revelou já ter comido **Al Caparras**, mas negou ter assistido ao filme **Al Capone**. [ex.s]

Considera-se também que **uso de *emoticons/emojis*** – presente no exemplo (33) – pode ser um recurso paralinguístico² para a identificação de um texto de notícia satírica, uma vez que dificilmente estes elementos apareceriam em uma notícia verídica.

- (33) Continuação do filme da Lava Jato vai se chamar “A Lei é para Todos kkkkkk 🤔🤔🤔”. [ex.s]

Enfim, as repetições lexicais são utilizadas para enfatizar uma determinada palavra na sentença, como no exemplo (34), ou no texto, como no exemplo (35). Em (34) o uso de “*Ciro*” em “*deixar de ser* *Ciro*” passa a ter um sentido figurado – remetendo não ao político *Ciro* Gomes, mas fazendo referência à pessoa que fica nervosa facilmente. Isso é reforçado em (35) na sentença “*Segundo* *Ciro*, *Ciro* ama *Ciro* mas o odeia por ter um pavio tão curto”.

- (34) **Ciro** dá um pescotapa em **Ciro** para ele deixar de ser **Ciro**. [ex.s]
- (35) O candidato a presidente **Ciro Gomes** deu um tapa no próprio pescoço depois de perder a cabeça mais uma vez. **Ciro** caiu na provocação do blogueiro do MBL Arthur do Val e deu um pescotapa no rapaz. O incidente aconteceu depois que Arthur lembrou as declarações de **Ciro**, que disse que sequestraria Lula em caso de condenação e que receberia Moro a tiros. Por sua vez, Arthur disse que não era Patrícia Pillar para apanhar de **Ciro**. A assessoria de **Ciro** disse que **Ciro** não bateu em **Ciro**. Amigos dizem que *Ciro* não se entende com **Ciro** porque é difícil ser **Ciro**. Segundo **Ciro**, **Ciro** ama **Ciro** mas o odeia por ter um pavio tão curto. **Ciro** divulgou nota negando **Ciro**. [ex.s]

5.1.2 Característica sintática

Assim como os aspectos lexicais e morfológicos, analisou-se a estrutura sintática de algumas sentenças com o propósito de encontrar sinais que apontam para uma notícia satírica. Um aspecto sintático observado nesta pesquisa é a **troca do sujeito**. O exemplo (36), retirado das notícias verdadeiras, o sujeito *Marcela Temer* executa a ação de *pular no lago para salvar o cachorro*. No entanto, em (37), referente satírico, o sujeito passa a ser o *cachorro* que *se joga no lago*, o que causa a estranheza ao leitor e produz o efeito de humor à sentença. Isso acontece por não ser muito verossímil a sequência sintática de um sujeito não-humano (como, *cachorro*), se jogar conscientemente (como, *por ter que conviver com Temer*) como acontece no exemplo da sentença verdadeira.

² Para Gavioli (2016, p. 247), “os emojis [são] códigos visuais paralinguísticos que dominaram as plataformas digitais para facilitar a comunicação digital, produzir sentido e conotar emoção durante as conversas.”

- (36) **Marcela Temer pula** em lago para salvar seu cachorro e afasta segurança que não ajudou. [ex.r]
- (37) **Cachorro** de Marcela **se jogou** no lago por ter que conviver com Temer. [ex.s]

5.1.3 Características semânticas

As características semânticas são características relevantes para a construção do sentido humorístico nas notícias satíricas, pois o sentido vai além do sentido literal de uma palavra ou expressão. O **meme**, por exemplo, reproduz ideias, informações e expressões frequentes nas redes sociais. Segundo Trevisan, Prá e Goethel (2016, p. 281), no meio digital “o termo meme passou a designar uma recente apropriação nas redes sociais, como a junção de imagens da cultura popular com frases que refletem pensamentos individuais ou coletivos, sejam eles com fins cômicos, políticos, etc.” Por ser um elemento da cultura digital popular, não se espera que aconteça em notícias verdadeiras. No exemplo (38) a expressão ‘*pedir música no Fantástico*’ significa que uma determinada ação que aconteceu várias vezes. A expressão se refere a uma prática do Fantástico, programa da Globo, em que jogadores de futebol podem pedir uma canção quando marcam três gols na mesma partida.

- (38) Temer, se roubar mais uma faixa presidencial, poderá **pedir música no Fantástico**. [ex.s]

Igualmente aos memes, as **expressões idiomáticas** ocorrem quando um conjunto de palavras possuem um sentido diferente daquele que as palavras teriam isoladamente. No exemplo (39), *pagar o pato* é utilizado no sentido de alguém levar a culpa por algo que não fez. O uso das expressões idiomáticas não acontece necessariamente em notícias satíricas e podem ocorrer em jornais de grande circulação, no entanto, é mais comum aparecerem em editoriais e artigos de opinião.

- (39) E até o pato da Fiesp deve **pagar o pato**. [ex.s]

Outro aspecto semântico capaz de auxiliar a identificação de uma notícia satírica é o da característica **palavras fora do contexto de domínio**. Em (40), como já abordado em na Seção 4.2.2, os termos *Alckmin* e *Lava-Jato* pertencem ao campo lexical *político*, já *drible* e *Tite* referem-se ao domínio do *futebol*. Desse modo, o item lexical *drible*, no léxico do futebol, significa o *movimento com a bola, buscando esquivar-se do adversário*. Porém, na sentença satírica é empregado figurativamente, dando o sentido de *tentativa de enganar, de fazer com que alguém seja enganado*.

- (40) **Alckmin** dá um **drible** na **Lava-Jato** e é convocado por **Tite**. [ex.s]

O mesmo acontece no exemplo (41), quando *Cirque du Soleil* e *malabarismo* são palavras referentes ao campo lexical *circense*, enquanto *Waack* é um jornalista brasileiro.

Observa-se, portanto, que *malabarismo*, no contexto do circense, significa *a técnica exercida pelo malabarista*. No entanto, a palavra, nesse exemplo, aparece aplicada figurativamente como *a versatilidade em contornar situações difíceis ou adversas*.

- (41) **Cirque du Soleil** abre seleção no Brasil após exposições de **malabarismo** em defesa de **Waack**. [ex.s]

Dessa forma, tanto *drible* quanto *malabarismo* se utilizam de uma **ambiguidade** para fazer uma relação entre um campo lexical e outro e, assim, gerar o humor da sátira.

Outro recurso sintático utilizado para resultar a sátira e o riso em notícias satíricas é a **quebra de expectativa** após *conjunções* em orações coordenadas e subordinadas. Comparando (42), notícia real, com (43), notícia satírica, é possível perceber que o efeito de humor é gerado pela quebra de expectativa após a conjunção *e*. O mesmo ocorre nos exemplos (44) e (45).

- (42) Fachin determina que Paulo Maluf comece a cumprir pena de mais de 7 anos de prisão. [ex.r]
- (43) STF determina a prisão de Maluf e presos começam a esconder pertences. [ex.s]
- (44) Joaquim Barbosa desiste da disputa presidencial. [ex.r]
- (45) Joaquim Barbosa desiste da presidência e mira o comando do Caldeirão do Huck. [ex.s]

Esse tipo de recurso evidencia a **incongruência** na sentença. Incongruência é um método, em que se utiliza muitas vezes do absurdo para expor algo inesperado. Como os eventos dos exemplos (43) e (45) são improváveis, a construção passa a ser satírica. Vale ressaltar que o uso do ‘e’ nas manchetes só é possível por conta do efeito implicatural de ‘e então’, por isso o humor acionado em 45).

5.1.4 Características estilísticas

As características estilísticas se referem ao grupo de aspectos linguísticos de figuras de linguagem utilizadas para a construção do efeito humorístico nas notícias satíricas. Muitas vezes um texto satírico se utiliza do absurdo para expor algo inesperado, como no exemplo (46).

- (46) O **papa** negou estar trabalhando pela **canonização** de **Lula**, como **afirmou o PT**. [ex.s]

O efeito de humor ocorre no exemplo acima a partir do absurdo provocado pela contraposição dos itens lexicais *papa* e *canonização* com o nome do ex-presidente *Lula* – na época condenado por corrupção pelo juiz Sérgio Moro.

Esse absurdo pode ser identificado muitas vezes pelo uso da *hipérbole*, sendo um indício de que o enunciado pertence a um texto satírico. Em (47), a sequência *para sempre*, destaca um modo eterno, infinito para o fim da *corrupção* no Brasil *chegada dos portugueses em 1500*, i.e, desde sua descoberta.

- (47) Com isso, o país se livra **para sempre** da **corrupção endêmica** que marcou sua vida política **desde a chegada dos portugueses em 1500**. [ex.s]

Além disso, notou-se que algumas construções das notícias satíricas se utilizam do *eufemismo* para criar a comicidade ao texto, como ocorre em (48).

- (48) Após começar **desempregando Dilma**, Temer atinge a marca de 13,7 milhões sem emprego.

No exemplo acima, o processo de impeachment que a ex-presidenta Dilma Rousseff sofreu em 2016 é retomado pela expressão *desempregar Dilma*. Neste caso, não houve uma demissão, mas um processo de impedimento de uma ex-presidente do Brasil, iniciado por uma briga política entre Dilma e o seu então vice, Michel Temer.

Ressalta-se também o uso da **metáfora** como um recurso linguístico nos textos das notícias satíricas. No exemplo (49), Gilmar Mendes, ministro do Supremo Tribunal Federal, foi responsável por soltar Paulo Preto, operador do PSDB e condenado pela Lava-Jato e o trecho *lembrou a ele de não se esquecer do casaquinho porque a noite estava fria* faz referência ao senso comum de que toda mãe diz para o filho levar um casaco. Esse sentido maternal na sentença é uma comparação ao ato de Gilmar Mendes soltar um filiado do PSDB³ ao cuidado que a mãe tem por um filho.

- (49) Depois de soltar Paulo Preto, Gilmar **lembrou a ele de não se esquecer do casaquinho porque a noite estava fria**. [ex.s]

A **personificação** também pode ser um sinal de uma notícia satírica. Tanto no exemplo (50) quanto no exemplo (51), objetos como *móvel* e *panela de pressão* passam a ter características humanas, como *dor* e *medo*. Apesar de também possuir características sintáticas – pois, *panela de pressão* é um complemento de *medo* e não de *Paulo* – esses casos são considerados aspectos lexicais por serem dependentes de outros itens lexicais para que ocorra a personificação.

- (50) Se ele bate com um dedinho em um móvel sem querer, o **móvel** é quem fica com **dor**. [ex.s]

- (51) A **panela de pressão** que tem **medo** de Paulo explodir. [ex.s]

³ No mesmo ano, o ministro do Supremo Tribunal Federal (STF), Gilmar Mendes, mandou arquivar uma investigação do então senador Aécio Neves, também filiado do PSDB. Disponível em: <<https://g1.globo.com/politica/noticia/2018/10/23/gilmar-mendes-arquiva-mais-um-inquerito-sobre-aecio-neves.ghtml>. Acesso em: 15 mai. 2022.>

Por fim, a **ironia** é um mecanismo que acontece pelo choque entre o significado literal do enunciado e o que se espera do falante. Deste modo, considera-se ironia uma característica linguística na sentença da notícia satírica quando há a oposição de sentido entre o que é pretendido e o que é realmente dito.

- (52) Crivella diz que **honestidade** de Garotinho é **tão real** quanto **Adão e Eva e Arca de Noé**. [ex.s]

No exemplo 52, o nome *honestidade* se liga ao adjetivo *real* por meio da comparação entre a **honestidade de Garotinho** e as *histórias bíblicas de Adão e Eva e da Arca de Noé*. É essa comparação que passa a indicar que o sentido da sentença é o oposto do que é dito: Garotinho, na verdade, é desonesto.

5.1.5 Resultados da análise das características linguísticas

Esta seção apresenta os resultados da análise das características linguísticas do *subcorpus* satírico. Assim, no Quadro 7 são apresentadas as características linguísticas, divididas em 4 características e subdivididas em 17 características específicas.

Quadro 7 – Características linguísticas das notícias satíricas

Características Linguísticas	
Características	Características Específicas
Características Lexicais	Palavras pejorativas (Pej)
	Vulgarismos (Vul)
	Autoria (Aut)
	Jogo de Palavras (Jpa)
	Emoticons/emojis (Emo)
	Repetição (Rep)
Característica Sintática	Troca do Sujeito (TSuj)
Característica Semântica	Expressões Idiomáticas (EId)
	Palavras fora do contexto de domínio (FDC)
	Ambiguidade (Amb)
	Meme (Meme)
	Quebra de Expectativa (QExp)
Características Estilísticas	Eufemismo (Euf)
	Metáfora (Met)
	Exagero (Exa)
	Personificação (Pers)
	Ironia (Ironia)

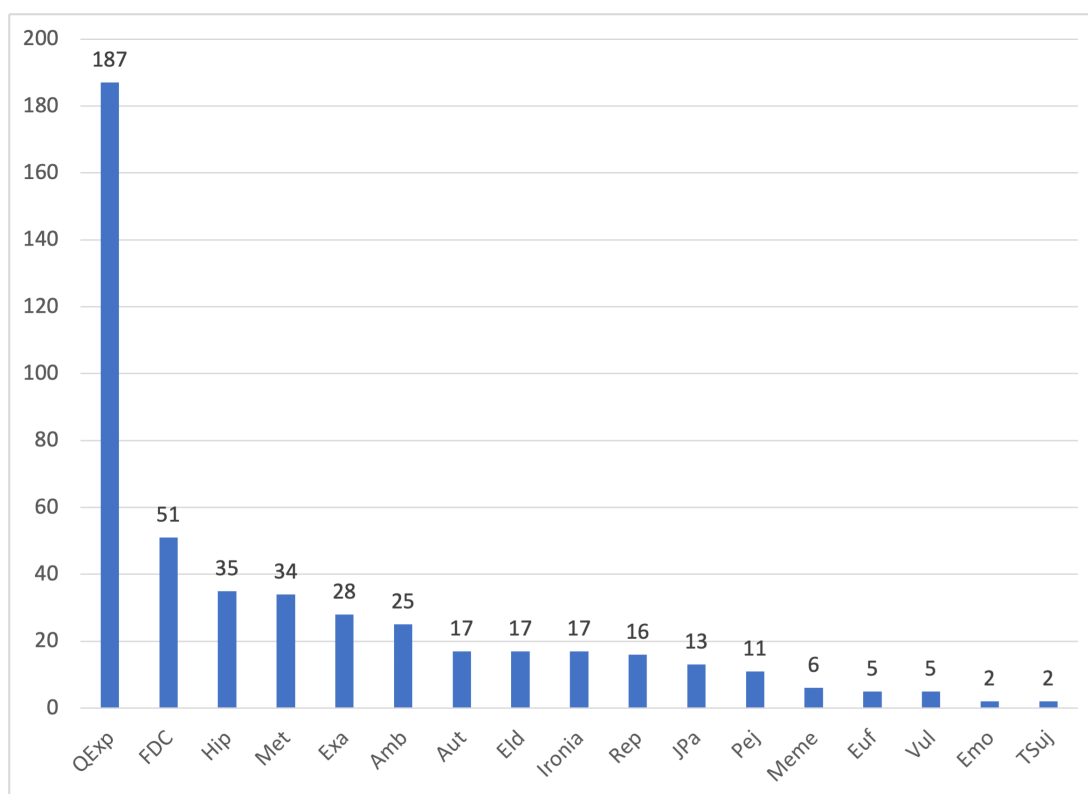
Fonte: Elaborada pela autora.

Na Figura 22, tem-se a disposição numérica dos atributos de cada característica linguística, por ordem de frequência, para facilitar a comparação. Identificou-se um total de 541 atributos linguísticos presentes nas sentenças satíricas, dos quais a quebra de expectativa (QExp) representa a maioria de ocorrências (15% do total). A seguir, vem as palavras fora do contexto do domínio (FDC) (4,09%). Nota-se que além dos dois atributos serem características semânticas, eles compartilham de, em alguns casos, necessitar de uma oposição de eventos e/ou assuntos diferentes para criar o sentido de humor.

(53) Lula não se apresenta e Ana Furtado é vista embarcando para Curitiba. [ex.s]

(54) O STF admitiu que se inspirou na cantora Anitta ao decidir adiar a decisão. [ex.s]

Figura 22 – Frequência dos atributos linguísticos



Fonte: Elaborada pela autora.

A Tabela 4 mostra que as características semânticas representam mais da metade (60,72%) do total das características que contêm nas sentenças das notícias satíricas. Este resultado demonstra que a partir dos recursos identificados manualmente, os recursos semânticos são os mais frequentes, enquanto os recursos sintáticos são menos evidentes (0,42%). Além disso, é possível afirmar que a sátira se constrói majoritariamente no nível semântico-pragmático da língua.

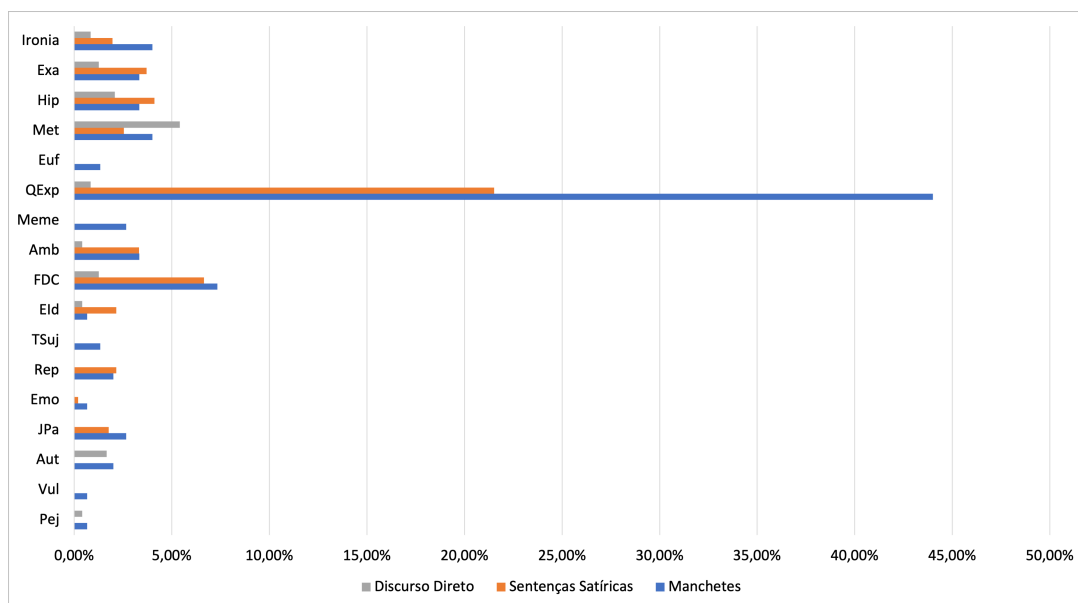
Tabela 4 – Frequência das características linguísticas

Características	Quantidade
Semânticas	286 (60,72%)
Estilísticas	119 (25,26%)
Lexicais	64 (13,58%)
Sintáticas	2 (0,42%)

Fonte: Elaborada pela autora.

Na Figura 23, é possível visualizar o gráfico do comparativo da disposição dos atributos linguísticos em cada tipo de sentença das notícias satíricas (manchetes, sentenças satíricas e discurso direto):

Figura 23 – Disposição das características linguísticas no *subcorpus* satírico



Fonte: Elaborada pela autora.

Vê-se que a quebra de expectativa (QExp) é substancialmente mais frequente em manchetes (44%) em relação às sentenças satíricas (21,52%) e as sentenças de discurso direto (0,83%). Uma suposição é que por meio desta característica, é possível contextualizar o leitor com o evento real, mas também ser humorística quando rompe com a expectativa ao trazer o evento satírico para o enunciado. Outra característica significativa é a palavra fora de contexto de domínio (FDC), sendo mais frequente nas manchetes (7,53%) quando comparadas às sentenças satíricas (6,65%) e discurso direto (1,25%). É possível notar ainda que a metáfora (Met) é mais frequente no discurso direto (5,41%) quando comparado às manchetes (4%) e sentenças satíricas (2,54%).

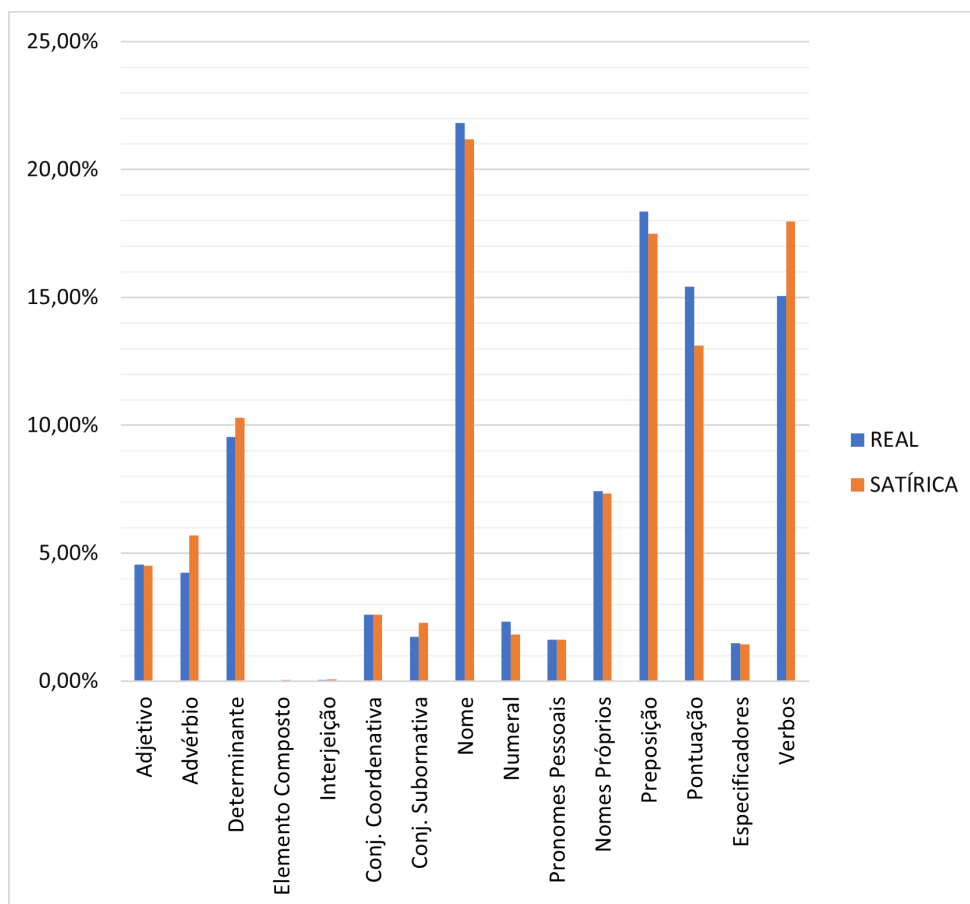
5.2 RESULTADOS OBTIDOS PELAS FERRAMENTAS DE PLN

5.2.1 Características extraídas do PALAVRAS

As informações morfossintáticas são provenientes do *parser* PALAVRAS (BICK, 2000). Além da anotação sintática, a ferramenta marca, para cada palavra, sua classe gramatical: adjetivo, advérbio, determinante, elemento composto, interjeição, conjunção coordenativa, conjunção subordinativa, nome, numeral, pronomes pessoais, nomes próprios, preposição, especificadores e verbos.

Com base nos dados extraídos do PALAVRAS, foi calculada a média entre a classe gramatical e o número total de palavras. Nota-se um balanceamento das classes gramaticais entre as notícias satíricas e as factuais. O **uso de advérbios** (5,69% nas notícias satíricas e 4,24% nas notícias reais), **determinantes** (10,30% nas notícias satíricas e 9,55% nas notícias reais) e **verbos** (17,98% nas notícias satíricas e 15,06%) ocorrem percentualmente mais nas notícias satíricas, enquanto as preposições (18,35% nas notícias reais e 17,50% nas notícias satíricas) e a pontuação (15,42% nas notícias reais e 13,12% notícias satíricas) são mais frequentes nas notícias verdadeiras, como mostra a Figura 24.

Figura 24 – Classes gramaticais extraídas pelo *parser* PALAVRAS (BICK, 2000)



Fonte: Elaborada pela autora.

Também foram analisados os tempos e pessoas verbais com o propósito de encontrar características específicas entre as notícias. Como é possível verificar na Tabela 5, não existe nenhum tempo verbal com maior predominância em relação às notícias, apenas uma maior porcentagem de infinitivo nas notícias satíricas (21,35%) quando comparado às notícias verdadeiras (16,45%). Uma possibilidade é que as notícias reais tendem a utilizar mais verbos auxiliares em relação às notícias satíricas, entretanto, o *parser* PALAVRAS não possui uma etiqueta específica para verbos auxiliares. A relação percentual foi calculada entre a frequência de cada tempo verbal pelo total de verbos anotados pelo *parser*.

Tabela 5 – Relação de tempo verbal entre notícias reais e satíricas

		Reais	Satíricas
Indicativo	Pretérito imperfeito	4,33%	3,39%
	Pretérito perfeito	21,77%	20,94%
	Presente	25,91%	24,60%
	Futuro	2,04%	2,89%
	Futuro do pretérito	1,83%	2,02%
Subjuntivo	Pretérito imperfeito	0,60%	0,48%
	Presente	2,18%	2,45%
Gerúndio		3,34%	5,03%
Infinitivo		16,45%	21,35%
Particípio		16,34%	13,08%

Fonte: Elaborada pela autora.

Sobre a análise das pessoas verbais, esperava-se que as notícias reais tivessem uma incidência maior de verbos na terceira pessoa do singular e do plural, pois como indica Tavares (1997, p. 130–131), “o texto jornalístico é caracterizado pela impessoalidade do sujeito”, uma vez que a “pessoa verbal que se reporta ao referente (aquele de quem se fala – “ele”, “eles”), [possibilita] que o texto seja mais objetivo”. A autora ainda salienta que o uso da primeira e da segunda pessoa não é o esperado para o texto jornalístico, porque tornam o texto mais subjetivo e pessoal. Assim, a partir dos dados descritos na Tabela 6, obtidos pelo PALAVRAS, observa-se que as notícias satíricas, que apesar de não serem baseadas apenas na realidade, possuem uma incidência maior de verbos na primeira e na terceira pessoa do singular, enquanto as notícias reais possuem mais verbos na primeira e terceira pessoa do plural.

Tabela 6 – Relação da pessoa verbal entre notícias reais e satíricas

	Reais	Satíricas
1ª Pessoa do Singular	14,31%	16,22%
2ª Pessoa do Singular	0,80%	0,26%
3ª Pessoa do Singular	56,64%	59,84%
1ª/3ª Pessoa do Singular	9,69%	12,23%
1ª Pessoa do Plural	3,29%	2,39%
3ª Pessoa do Plural	10,79%	8,77%

Fonte: Elaborada pela autora.

Vale ressaltar que essa análise não considerou a diferenças das ocorrências verbais nos 1) discursos diretos/indiretos, como nos exemplos (55) e (56); 2) relação do uso na 3ª pessoa do singular ao se referir ao autor ou veículo, como no exemplo (57):

- (55) “Se as pessoas já podem andar por aí portando meus CDs e DVDs, materiais bélicos de alta periculosidade, por que não um fuzil que dispara 100 tiros por minuto?”, perguntou. [ex.s]
- (56) Um vidente disse que a queda de Cunha e o Ovomaltine no McDonalds são sinais evidentes de que o fim está mesmo próximo. [ex.s]
- (57) O Sensacionalista apurou que o jovem se defende dizendo que, há 80 anos, seu bisavô teria dirigido a palavra a um negro. [ex.s]

5.2.2 Características extraídas pelo LIWC

O LIWC foi aplicado para a obtenção das medidas linguísticas para a análise dos textos das notícias reais e satíricas. Como já foi apresentado no Capítulo 4, a função da ferramenta é ler cada palavra em um texto (ou grupo de palavras) e combinar essa palavra com dicionários de referência carregados na memória do programa. Yang, Mukherjee e Dragut (2017) enfatizaram a importância das características psicológicas, uma vez que as notícias reais tendem a ser mais conservadoras e as satíricas mais agressivas. Salas-Zárate et al. (2017) apontaram cinco características psicológicas significativas da identificação de uma notícia satírica: processo social, processo afetivo, emoções positivas, processo cognitivo e certeza.

A Tabela 7 indica, por exemplo, que 5,046% das palavras presentes nas notícias satíricas e 4,754% das palavras existentes nas notícias reais estão incluídas na categoria *causa*, ou então, 4,051% das palavras nas notícias reais e 3,506% se enquadram na categoria **trabalho**. Observa-se ainda que não há nenhuma categoria completamente divergente entre as notícias reais e satíricas. Os textos satíricos têm níveis mais altos de palavras relacionadas à *tentativa*, *causa*, *certeza*, mas níveis mais baixos de palavras de *trabalho*, o que pode indicar que por pertencer ao domínio político, as notícias verdadeiras abordam

mais assunto sobre o trabalho de cargos políticos ou jurídicos, enquanto as notícias reais podem estar mais associadas a informar sobre o aumento ou diminuição de emprego. Além disso, como destaca Salas-Zárate et al. (2017), os níveis mais altos de palavras de *certeza* (como *nunca* ou *sempre*) nas notícias satíricas (3,156%) em relação às reais (2,912%), o que pode estar associados à hipérbole (já discutido na Seção 5.1.4). Ressalta-se também que 8,510% das palavras das notícias satíricas e 7,364% das palavras das notícias reais se enquadra na categoria *tentativa* (como *pode ser*, *talvez*, *acho*), o que indica que os textos satíricos se utilizam mais de modalizações, produzindo um efeito de sentido e expressa uma intenção de possibilidade e incerteza.

Tabela 7 – Relações psicolinguísticas a partir dos dados extraídos do LIWC

Categorias	Exemplos	Reais	Satíricas
alcance (<i>achieve</i>)	vitória, sucesso, melhor	5,709%	5,805%
afeto (<i>affect</i>)	feliz, chorou	2,720%	3,927%
raiva (<i>anger</i>)	odiar, matar, irritado	1,272%	1,493%
ansiedade (<i>anx</i>)	preocupado, com medo	0,597%	0,934%
causa (<i>cause</i>)	porque, efeito	4,754%	5,046%
certeza (<i>certain</i>)	sempre, nunca	2,912%	3,156%
morte (<i>death</i>)	enterrar, caixão, matar	0,332%	0,396%
discordância (<i>discrep</i>)	deveria, seria	5,170%	5,814%
família (<i>family</i>)	filha, pai, tia	0,595%	0,518%
sentir (<i>feel</i>)	sente, toque	1,659%	1,987%
amizade (<i>friend</i>)	amiga, vizinho	0,554%	0,587%
ouvir (<i>hear</i>)	ouvir, escutar	1,802%	2,231%
lar (<i>home</i>)	cozinha, senhorio	0,915%	0,951%
inibição (<i>inhib</i>)	bloquear, restringir	6,050%	6,249%
intuição (<i>insight</i>)	pense, saiba	5,803%	5,869%
lazer (<i>leisure</i>)	cozinhar, conversar, filme	1,306%	1,987%
dinheiro (<i>money</i>)	auditoria, dinheiro, dever	2,868%	2,841%
religião (<i>relig</i>)	altar, igreja	0,812%	0,938%
tristeza (<i>sadness</i>)	chorando, pesar, triste	0,714%	0,861%
ver (<i>see</i>)	visão, viu, visto	1,174%	1,040%
palavrão (<i>swear</i>)	foda, porra, merda	7,021%	7,072%
tentativa (<i>tentat</i>)	pode ser, talvez, acho	7,364%	8,510%
trabalho (<i>work</i>)	emprego, xerox	4,051%	3,506%

Fonte: Elaborada pela autora.

Tabela 8 – Frequência de emoções das notícias com base no dicionário do LIWC

Emoções	Real	Satírica
positiva	0,109	0,392
negativa	0,024	0,174
neutra	0,003	0,072

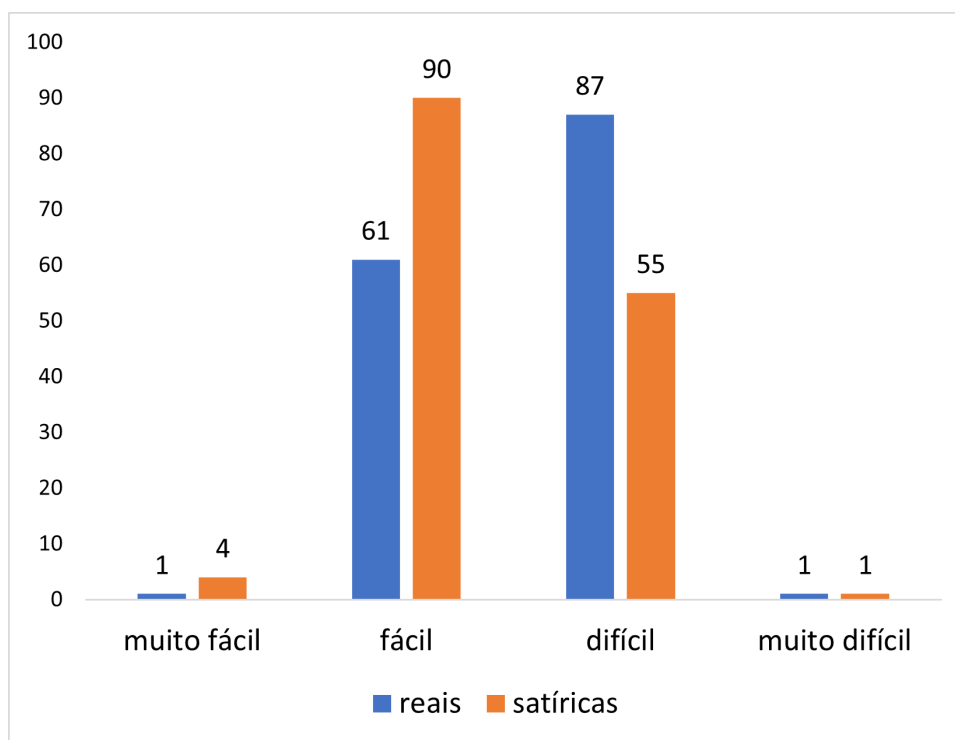
Fonte: Elaborada pela autora.

Além da categoria afeto ser evidente nas notícias satíricas, validando que o texto satírico é mais emocional, a partir dos dados apresentados na Tabela 8, é possível observar que as notícias satíricas empregam consideravelmente mais palavras positivas (0,392) e negativas (0,174) do que as notícias reais. Desse modo, é possível afirmar que o uso de emoções (positivas, negativas e neutras) são mais evidentes em notícias satíricas. É interessante pensar que o uso da emoção pode estar ligado ao fato de (1) notícias reais tendem a utilizar uma linguagem mais imparcial em relação às satíricas e (2) o uso excessivo de palavras positivas pode, na verdade, mostrar o uso de recursos como a ironia, que se utiliza de palavras ou sentenças com a polaridade oposta ao sentido literal, como também apontou Salas-Zárate et al. (2017).

5.2.3 Características extraídas pelo NILC-Matrix

5.2.3.1 Análise Índice Flesch

Como já discutido (ver Seção 4.3.2.1), o Índice Flesch é uma medida utilizada para indicar a dificuldade de compreensão durante a leitura. A Figura 25 apresenta a estatística da leiturabilidade das notícias satíricas e verdadeiras de acordo com o Índice Flesch Brasileiro (MARTINS et al., 1996). Para a análise, foi aplicado o cálculo do Índice Flesch para o português nas 150 notícias reais e nas 150 notícias satíricas. Conforme a descrição do Quadro 8 e as informações apresentadas na terceira coluna do gráfico, as notícias verdadeiras são mais difíceis (87) em relação às satíricas (55). Ainda, em comparação, a primeira coluna do gráfico mostra que as notícias satíricas (4) são muito mais fáceis do que as notícias verdadeiras (1). A segunda coluna indica uma facilidade de leiturabilidade muito maior nos textos satíricos (90) em contraste com os textos verídicos (61).

Figura 25 – Estatística de leitura do *corpus* conforme o Índice Flesch Brasileiro

Fonte: Elaborada pela autora.

Quadro 8 – Leiturabilidade do *corpus* de acordo com o Índice Flesch Brasileiro

Escore	Nível de Complexidade	Grau Escolar
100-75	Muito Fácil	1º a 5º ano
75-50	Fácil	6º a 9º ano
50-25	Difícil	Ensino Médio
25-00	Muito Difícil	Ensino Superior

Fonte: (FINATTO et al., 2016).

5.2.3.2 Análise NILC-Matrix

Semelhante ao trabalho de Levi et al. (2019), que aplicou as medidas do Coh-Matrix (MCNAMARA et al., 2010) para testar a hipótese de que as métricas de coerência textual podem ser úteis para capturar aspectos semelhantes do relacionamento semântico entre as sentenças de uma notícia, esta seção analisa as principais diferenças de inteligibilidade entre notícias satíricas e reais através do NILC-Matrix (LEAL, 2021). Para a análise da inteligibilidade das notícias satíricas e reais, foram selecionadas apenas as medidas com resultados mais expressivos, separadas nas 7 categorias seguintes: (i) medidas descritivas; (ii) coesão referencial; (iii) diversidade lexical; (iv) leiturabilidade; (v) complexidade sintática; (vi) informações morfosintáticas e (vii) informações semânticas. Os resultados são apresentados na Tabela 9.

Tabela 9 – Principais resultados extraídos do NILC-Metrix

CATEGORIA	MÉTRICAS	REAIS	SATÍRICAS
MEDIDAS DESCRITIVAS	Frequência de palavras de conteúdo	611.825,41	490.304,19
	Número de palavras	594,08	156,44
	Número de sentenças	32,02	8,36
	Palavras por sentença	19,35	19,49
COESÃO REFERENCIAL	Pronome anafórico do caso reto	0,27	0,19
	Pronome demonstrativo anafórico	0,30	0,21
	Referência anafórica	1,55	1,05
	Referência anafórica (adjc.)	0,33	0,24
	Sobreposição de argumentos	0,70	0,97
	Sobreposição de argumentos (adjc.)	0,84	1,06
	Sobreposição de radical palavras	0,93	1,38
	Sobreposição de radical de palavras (adjc.)	1,14	1,53
DIVERSIDADE LEXICAL	TTR	0,73	0,73
	Diversidade de preposições	0,28	0,49
	Diversidade de pronomes	0,55	0,76
	Diversidade de pontuação	0,08	0,17
	Diversidade de pronomes relativos	0,19	0,43
	Diversidade de verbos	0,77	0,88
LEITURABILIDADE	Brunet	12,19	10,43
	Honore	889,01	920,29
COMPLEXIDADE SINTÁTICA	Preposições por sentença	1,53	2,16
	Média de orações por sentenças	0,96	1,83
	Distância de dependências	45,36	43,94
	Proporção de conj. coordenativas	0,27	0,35
	Proporção de conj. subordinativas	0,13	0,28
	Tamanho médio dos SN	5,25	4,89
INFORMAÇÕES MORFOSSINTÁTICAS	Pronomes em 1ª Pessoa	0,12	0,11
	Pronomes em 3ª Pessoa	0,71	0,49
	Verbos flexionados	0,25	0,38
	Verbos não flexionados	0,15	0,25
INFORMAÇÕES SEMÂNTICAS	Ambiguidade	12,84	13,89

Fonte: Elaborada pela autora.

Como esperado, os índices das medidas descritivas – frequência de palavras de conteúdo, números de palavras, números de sentenças – têm valores mais elevados em notícias verdadeiras, pois geralmente são textos mais longos e, conseqüentemente, mais complexos. Em relação aos mecanismos coesivos de referenciação textual, as medidas de referência anafórica abaixo se mostram mais presentes em textos reais:

- **Pronome anafórico do caso reto:** Relaciona-se à média de candidatos a referente, na sentença anterior, por pronome anafórico do caso reto. São considerados apenas os pronomes anafóricos do caso reto: *ele*, *ela*, *eles*, *elas*. Segundo Leal (2021), o

referente do pronome anafórico é procurado na sentença adjacente anterior. Em (58), por exemplo, o pronome *ele* é um referente para *Paulo Vieira de Souza* e *Paulo Preto*

(58) O ministro do Supremo Tribunal Federal (STF) Gilmar Mendes mandou soltar, no início da noite desta quarta-feira, o ex-diretor da empresa paulista de Desenvolvimento Rodoviário (Dersa) **Paulo Vieira de Souza**, conhecido como **Paulo Preto**, preso nesta manhã pela Polícia Federal (PF). **Ele** é suspeito de ser o operador do PSDB no esquema de propina investigado pela Operação Lava Jato e é acusado de desvios de 7,7 milhões de reais da Dersa. [ex.r]

- **Pronome demonstrativo anafórico:** Relaciona-se à média de candidatos a referente, na sentença anterior, por pronome demonstrativo anafórico. Segundo Leal (2021), a referência anafórica é a relação entre um pronome e o termo anterior que ele substitui, assim, o pronome é a anáfora e o nome que ele substitui é o referente. Em (59), o pronome *desse* é um referente para *grupo político apelidado de “partido da Alerj”*.

(59) Picciani, Albertassi e Melo são expoentes do **grupo político apelidado de “partido da Alerj”**, os velhos caciques que mandam no Legislativo fluminense há mais de 20 anos e que também possuem forte influência no Executivo e no Judiciário local. **Desse** grupo saiu o ex-governador Sérgio Cabral (PMDB), que também foi presidente da Assembleia Legislativa. [ex.r]

- **Referentes anafóricos e referentes anafóricos adjacentes (adjc.)**⁴ Relaciona-se Média das proporções de candidatos a referentes nas 5 sentenças anteriores em relação aos pronomes anafóricos das sentenças ou sentenças adjacentes. Em (60), há 2 pronomes anafóricos: “ela” tem apenas 1 candidato a referente na sentença anterior (Andrea), enquanto ‘ele’ tem 2 candidatos a referente na sentença anterior (imóvel, Joesley Batista).

(60) Minha irmã, **Andrea**, ofereceu o **imóvel** a alguns empresários, inclusive ao senhor **Joesley Batista**. **Ela** teve com **ele**, em toda a sua vida, um único encontro, a meu pedido, motivado por esse assunto familiar que nada teve a ver com política. [ex.r]

Desse modo, por essas medidas se referirem a um recurso coesivo que busca a manutenção de sentidos apresentados anteriormente, quanto maior a métrica, maior a complexidade textual. Pode-se, portanto, verificar que o tamanho dos textos também

⁴ Segundo Leal (2021), sentenças adjacentes são todas as possíveis combinações de sentenças do texto em uma determinada sequência de sentenças.

influencia diretamente na quantidade de dêiticos. Assim, quanto mais texto, mais vezes é preciso repetir o que já foi dito.

Identificou-se também que em relação aos resultados das métricas de diversidade textual, observa-se que o valor do TTR é idêntico entre as notícias reais e as notícias satíricas. No entanto, quando as classes gramaticais são analisadas separadamente, como a diversidade de preposições, pronomes, pontuações, pronomes relativos e verbos, os índices são maiores nas notícias satíricas. Apesar de não estar clara a relação da métrica de diversidade dessas classes com a complexidade textual, de acordo com Leal (2021), supõe-se que, quanto maior métrica, maior a complexidade. As medidas de leiturabilidade, em contrapartida, são discordantes entre as estatísticas de Brunet e Honoré. Enquanto o Índice de Brunet, que é uma forma de TTR menos sensível ao tamanho do texto, elevando o número de *types* à constante -0,165 e depois eleva-se o número de tokens a esse resultado, é maior nas notícias reais (12,19) em relação às satíricas (10,43); o Índice de Honoré, um tipo de TTR que leva em consideração, além da quantidade de *types* e *tokens*, a quantidade de *hapax legomena*⁵, é maior nas notícias satíricas (920,29) comparadas às notícias reais (889,01).

As medidas de sobreposição de argumentos nas sentenças e nas sentenças adjacentes, relativas à quantidade média de referentes que se repetem nos pares de sentenças das notícias, indicando se a formação de uma cadeia de correferência é facilitada ou não e, sobreposição de radical de palavras que se repetem nos pares de sentenças e sentenças adjacentes, possuem maior valor nas notícias satíricas. Portanto, considerando que a repetição de referentes é um recurso de simplificação, quanto maior a métrica, menor a complexidade textual – com exceção em textos constituídos de uma única sentença.

Do ponto de vista da complexidade sintática, a respeito da distância de dependência e do tamanho médio dos sintagmas nominais (SN), as notícias reais têm valores maiores quando comparados com as notícias satíricas. O cálculo que mede a distância de dependência utiliza uma árvore de dependências sintáticas, sendo que cada relação de dependência está associada a uma distância entre as palavras na superfície textual e, desse modo, quanto maiores as distâncias de dependência, maior a complexidade da notícia. Quanto ao tamanho médio do SN e considerando que eles são constituintes de uma oração em que o núcleo é um nome e os demais integrantes, não obrigatórios, são determinantes, adjetivos e outros modificadores nominais, quanto maior o resultado, i.e., quanto maior o tamanho médio dos sintagmas nominais, maior a complexidade textual da notícia.

Com relação ainda às medidas de complexidade sintática, a média de orações por sentença e média de preposições por sentença são maiores nas notícias satíricas se comparadas com as notícias reais. Logo, quanto maior o número de orações por sentença, maior a complexidade. Similarmente, uma vez que as preposições introduzem argumentos

⁵ Um *legomenon hapax* é uma palavra ou expressão que ocorre apenas uma vez dentro de um contexto; seja no registro escrito de uma língua inteira, nas obras de um autor, ou em um único texto.

verbais e adjuntos modificadores, isso tende a aumentar a complexidade textual da notícia. Observa-se ainda que tanto a proporção de conjunções coordenativas quanto subordinativas são maiores nas notícias satíricas. Para Leal (2021), as conjunções coordenativas parecem ser índices de estruturas mais complexas que as estruturas simples, porém menos complexas que as estruturas com subordinação. Desse modo, as conjunções subordinativas, por introduzirem orações subordinadas, indicam estruturas mais complexas que as conjunções coordenadas e, portanto, quanto maior o resultado, maior a complexidade do texto. É interessante que a característica linguística mais relevante encontrada nas notícias satíricas é a quebra de expectativa, que consiste basicamente na ruptura do sentido inicial da sentença, como nos exemplos abaixo:

- (61) PF pede transferência de Lula e 31% dos brasileiros sugerem Palácio do Planalto. [ex.s]
- (62) Após deixar triplex, MTST quer ocupar sítio em Atibaia **mas** ninguém se voluntariou. [ex.s]
- (63) Estação espacial chinesa não caiu em SP **porque** Correios seguraram em Curitiba. [ex.s]
- (64) Temer explicou **ainda que** poderá utilizar o mesmo layout das urnas de 2014, como forma de cortar gastos. [ex.s]

Por fim, nota-se que a ambiguidade é mais evidente em notícias satíricas. Esse resultado se mostra relevante, pois, como já foi abordado neste capítulo e no capítulo anterior, a ambiguidade é uma característica da construção da sátira e da paródia, uma vez que seu uso intencional causa no leitor uma confusão de sentido, podendo gerar um efeito de humor no texto. Contudo, para a compreensão do sentido ambíguo, é necessário o conhecimento extralinguístico daquilo que se lê e, dessa forma, quanto mais sentidos um texto tiver, maior será o esforço requerido do leitor para a desambiguação.

5.3 ANÁLISE DAS INTERPRETAÇÕES DAS SENTENÇAS SATÍRICAS

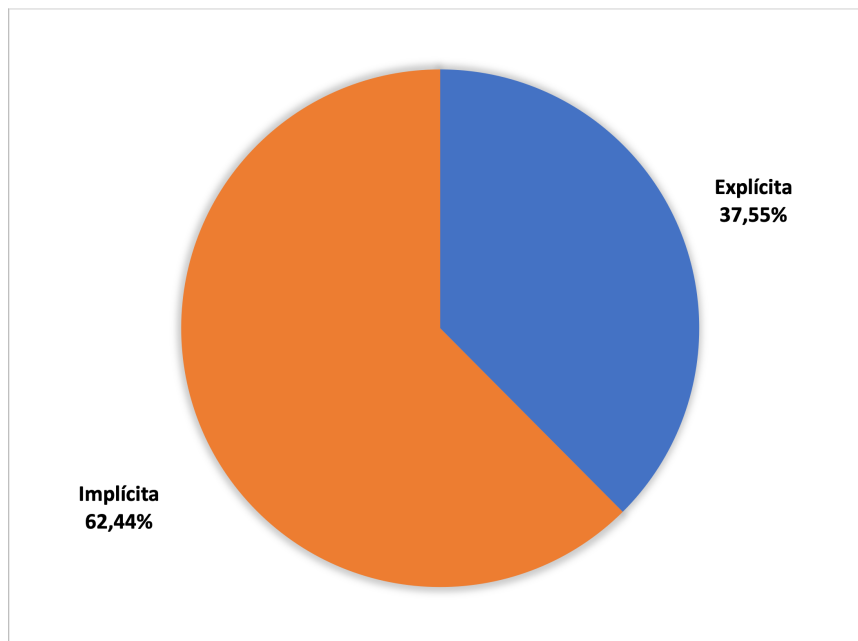
Como apresentado no capítulo anterior, uma sentença satírica pode ser classificada em duas categorias:

- **Explícita:** quando é possível compreender o efeito satírico a partir de sinais linguísticos no enunciado.
- **Implícita:** quando a sátira é entendida somente com o contexto extralinguístico.

Dessa forma, as sentenças das notícias satíricas (manchetes, sentenças satíricas e discurso direto) foram classificadas conforme suas interpretações. Assim, segundo a Figura

26 , das 900 sentenças (total das manchetes, sentenças satíricas e discurso direto satírico), 62,44% são interpretações implícitas e 37,55% são interpretações explícitas.

Figura 26 – Porcentagem de interpretação implícita e explícita



Fonte: Elaborada pela autora.

Na Tabela 10, descrevem-se os valores referentes à interpretação de cada tipo de sentença das notícias satíricas:

Tabela 10 – Total de sentenças classificadas como implícita e explícita

Sentenças	Qtd. sentenças	Explícitas	Implícitas
Manchetes	150	98	52
Satíricas	511	208	303
Discurso Direto	239	32	207
Total	900	306	562

Fonte: Elaborada pela autora.

Ao detalhar a interpretação dessas sentenças, na Tabela 11, tem-se a distribuição da porcentagem de valores referentes à interpretação explícita e implícita em relação à quantidade do total de sentenças e referente à quantidade de sentenças de cada tipo (manchetes, satíricas e discurso direto).

Tabela 11 – Interpretação implícita e explícita das sentenças das notícias satíricas

Sentenças	Qtd. sentenças totais		Qtd. tipo de sentenças	
	Explícitas	Implícitas	Explícitas	Implícitas
Manchetes	10,88%	5,88%	65,33%	35,33%
Satíricas	23,11%	33,66%	40,70%	59,29%
Discurso Direto	3,55%	23%	13,38%	86,61%

Fonte: Elaborada pela autora.

Observa-se que a interpretação implícita e explícita das sentenças das notícias satíricas em relação às 900 sentenças das notícias satíricas (com exceção das classificadas como sentenças reais). Assim, 303 (33,66%) das 900 sentenças totais das notícias são implícitas, ao passo que 208 (23,11%) são explícitas e 207 (23%) das 239 sentenças de discurso direto das notícias satíricas são implícitas enquanto 32 (3,55%), são explícitas. Em contrapartida, 96 (10,88%) das manchetes foram classificadas como explícitas e 52 (5,77%) como implícitas.

Ainda é possível verificar que a porcentagem da relação das interpretações explícitas e implícitas entre cada tipo de sentença das notícias satíricas. Assim, das 150 manchetes satíricas, 98 (65,33%) são classificadas como explícitas, enquanto 52 (34,66%) são implícitas. Em relação às 511 sentenças classificadas como satíricas, 303 (59,29%) são implícitas e 208 (40,70%) são explícitas. Por último, quanto aos discursos diretos, 207 (86,61%) das 239 sentenças classificadas como discurso direto são implícitas à medida que 32 (13,38%) são explícitas.

5.3.1 Anotação de manchetes implícitas e explícitas

Para validar a classificação proposta na subseção anterior, foi feita uma anotação com três anotadores das 150 manchetes satíricas, em que deveriam classificar a sentença em explícita e implícita. A partir da anotação, calcula-se o acordo inter-anotadores, sendo o coeficiente *kappa* o mais conhecido (PUSTEJOVSKY; STUBBS, 2012). O *kappa* é uma forma de medir o grau de concordância entre dois (*kappa* de Cohen) ou três ou mais avaliadores (*kappa* de Fleiss) quando os avaliadores estão atribuindo classificações categóricas a um conjunto de itens. Por esta pesquisa considerar a anotação de três anotadores, a verificação da concordância inter-anotador da anotação será através do *kappa* de Fleiss, onde a equação de base para o cálculo do seu coeficiente é:

$$\kappa = \frac{P - P_e}{1 - P_e}$$

Na equação acima, P representa o acordo real e P_e representa acordo esperado. Para melhor representar os valores do anotador para o *kappa* de Fleiss, desenha-se uma tabela, a qual possui um eixo para as categorias possíveis que um anotador pode atribuir

seus valores e outro eixo para cada um dos anotadores. Na Tabela 12, vê-se as categorias no topo, os anotadores ao lado e o conteúdo de cada célula representa quantas vezes o anotador atribuiu essa etiqueta ao documento.

Tabela 12 – Categorias e anotadores para cálculo do coeficiente *kappa*

	Explícita	Implícita	P_i
Anotador A	75	75	0,4966
Anotador B	96	54	0,5361
Anotador C	71	79	0,4980
P_c	0,5377	0,4622	

Fonte: Elaborada pela autora.

Segundo Pustejovsky e Stubbs (2012, p. 129) para medir o *kappa*, inicialmente, deve-se calcular quantas atribuições foram dadas proporcionalmente para cada categoria. Isso é representado por P_c , sendo c a representação da categoria avaliada. O cálculo é realizado pela soma dos valores em sua coluna, dividido pelo número de anotadores vezes o número de anotações por cada anotador, onde A representa o número de anotadores, a número de anotação por anotadores e i a representação do atual anotador. Esse cálculo é representado pela seguinte equação:

$$P_c = \frac{1}{Aa} \sum_{i=1}^A a_{ic}, \quad 1 = \frac{1}{a} \sum_{c=1}^k a_{ic}$$

Então, se a equação acima for aplicada à categoria explícita, obtém-se a seguinte equação:

$$P_{explícita} = (75 + 96 + 71) / (3 \times 150)$$

O mesmo cálculo foi realizado na categoria implícita. O resultado é representado na última linha na Tabela 12.

Seguindo, foi calculado o P_i , i.e., o acordo de cada anotador com os outros anotadores em comparação a todos os valores de concordância possíveis. Isso significa que para cada linha, soma-se os quadrados dos valores de cada coluna, tendo o resultado pelo número de anotações totais por cada anotador. O cálculo é representado pela equação abaixo:

$$P_i = \frac{(\sum_{c=1}^k a_{ic}^2) - (a)}{a(a-1)}$$

Então, ao realizar a medida P_i para o Anotador A, é possível obter o resultado do **Anotador A** através do cálculo abaixo. Assim, ao calcular também o P_i para os **Anotadores B** e **C**, tem-se o resultado na quarta coluna da Tabela 12.

$$P(\text{AnotadorA}) = ((75^2 + 75^2) - 150) / 150(1501)$$

Em seguida, a partir dos resultados de P_c e de P_i , calculou-se o valor de P , que na equação original de Fleiss é a média dos valores de P_i . Então, para isso foram somados os valores da coluna P_i dividido pelo número de anotadores, resultando o valor 0,5102. Logo, calculado o P da equação, o próximo passo foi determinar o valor de P_e , sendo a soma dos quadrados dos valores de P_c , em que se obteve o valor 0,5028.

Finalmente, com base nos resultados, mediu-se o *kappa* da anotação. Como resultado, a anotação atingiu o valor 0,014 e, como mostra o Quadro 9, é um valor de concordância fraca.

Quadro 9 – Valores de interpretação do coeficiente *kappa*

<i>Kappa</i>	Concordância
<0	Ruim
0,01-0,20	Fraca
0,21-0,40	Sofrível
0,41-0,60	Regular
0,61-0,80	Boa
0,81-0,99	Ótima
1	Perfeita

Fonte: Elaborada pela autora.

A concordância baixa pode estar relacionada à dificuldade do anotador identificar uma sentença explícita ou implícita. Além disso, eventos reais ou personagens presente nas notícias que são desconhecidos pelo anotador também pode dificultar a compreensão do que é explícito ou implícito no texto. Desse modo, no exemplo 65, para o anotador que reconhece que “Ana Furtado”, é provável que entenda como um sinal explícito na sentença; já o anotador que não consegue captar que se trata de uma pessoa que não pertence ao domínio político, logo, é possível que entenda como um sinal implícito. Entende-se, portanto, que a identificação de uma sentença satírica é uma tarefa que exige não só a interpretação da intenção da sátira, mas pode envolver também o compartilhamento da cultura, de uma forma de humor e de conhecimento de mundo do leitor.

(65) Lula não se apresenta e **Ana Furtado** é vista embarcando para Curitiba. [ex.s]

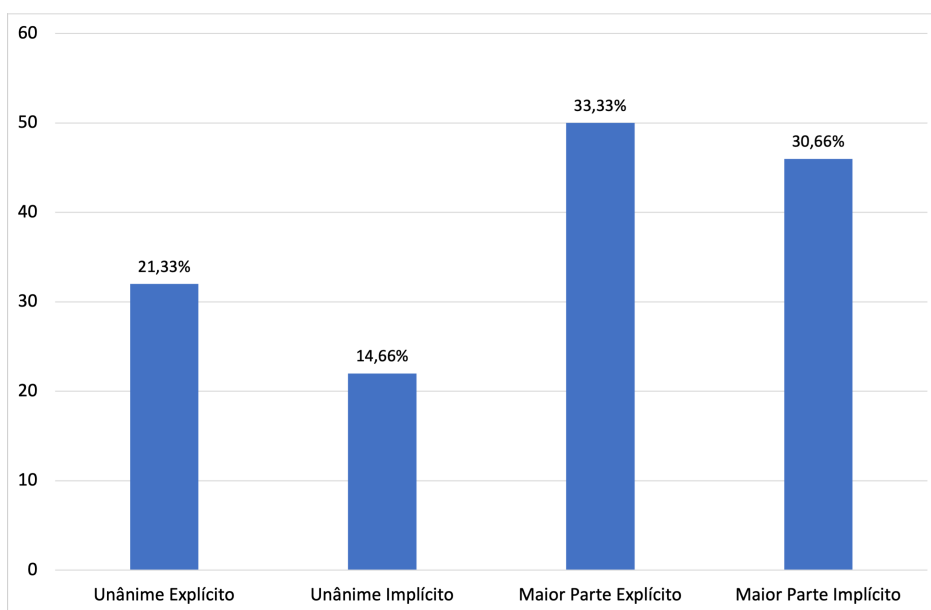
Contudo, para Pustejovsky e Stubbs (2012), a interpretação do cálculo *kappa* depende muito da complexidade e da objetividade do conteúdo que está sendo anotado. Os autores sugerem que em tarefas que envolvam anotações, como *part-of-speech*, espera-se que a anotação atinja uma pontuação κ próxima a 1 devido ao modo como são definidas as classificações, termos e teorias linguísticas. Por outro lado, em tarefas de anotação que exigem uma interpretação maior do anotador, como no caso da das interpretações implícitas e explícitas, geralmente, não se atinge um padrão alto de anotação. Por isso, é importante considerar que as sentenças anotadas estão estritamente ligadas ao contexto

da notícia e também ao próprio conhecimento extralinguístico do anotador, tornando a tarefa de medir o IAA mais complexa.

Além disso, vale destacar que o valor κ , descrito acima, mede apenas a concordância entre os anotadores, desconsiderando quais sentenças foram anotadas unanimemente pelos anotadores. Desse modo, em muitos casos na anotação, observou-se que cada anotador classificou diferentes manchetes como explícitas e implícitas, mas outras houve concordância entre a classificação de sua interpretação, o que se faz considerar que existem determinadas sentenças que são mais identificáveis, em relação ao seu caráter satírico, que outras.

Assim, como mostra a Figura 27, os anotadores concordaram de modo unânime que 32 (22,33%) das 150 manchetes são explícitas, enquanto 22 (14,66%) foram anotadas pelos três anotadores como implícitas. No entanto, 50 (33,33%) manchetes tiveram 2 anotadores que marcaram como sendo explícita e 1 como implícita e, por fim, 46 (30,66%) das 150 foram anotadas por 2 anotadores como implícita e 1 como explícita.

Figura 27 – Concordância entre os anotadores



Fonte: Elaborada pela autora.

Em uma análise mais detalhada das 33 manchetes unânimes explícitas e das 22 unânimes implícitas, observou-se que existem algumas semelhanças a serem destacadas. Os exemplos (66) e (67) são duas manchetes classificadas pelos três anotadores como implícita. Nelas, a construção da sátira nas sentenças é muito sutil, sendo necessário alguns cálculos cognitivos e contextuais para retomar os eventos (E) das quais as duas notícias abordam e compreendê-las como satírica.

(66) Lula diz que frase na série de Padilha não é sua, é de um amigo. [ex.s]

(67) Ana Furtado substituirá Huck como candidata à presidência. [ex.s]

No exemplo (66), é necessário que o leitor tenha o conhecimento de que (E1) existe uma série escrita por José Padilha; (E2) a série escrita por Padilha, chamada ‘O Mecanismo’, é sobre uma investigação sobre corrupção envolvendo estatais e empreiteiras se torna um dos maiores escândalos políticos do Brasil, fazendo referência à Operação Lava-Jato; (E3) na série, existe uma personagem, João Higino, inspirada no ex-presidente Lula; (E4) João Higino diz que é preciso “construir um acordo nacional para estancar a sangria”; (E5) saber que a frase foi dita por Romero Jucá a Sérgio Machado em como paralisar as investigações da Lava-Jato contra membros do PMDB e do governo Temer; (E6) sendo que Romero Jucá foi um dos responsáveis pelo impeachment da ex-presidenta Dilma Rousseff em 2016; (E7) ainda em 2016, o ex-presidente Lula foi acusado pela Operação Lava-Jato, de ser o verdadeiro proprietário do sítio, que estava em nome de seu amigo, Fernando Bittar. Assim, tem-se a seguinte sequência de eventos:

$$E1 \rightarrow E2 \rightarrow E3 \rightarrow E4 \rightarrow E5 \rightarrow E6 \rightarrow E7$$

Em (67), é necessário (E1) saber que Ana Furtado é uma apresentadora da Rede Globo; (E2) virou meme por ser chamada frequentemente para substituir outros apresentadores da emissora; (E3) Luciano Huck também é um apresentador de televisão; (E4) em 2018, após afirmar que se candidataria ao cargo presidente da República, Luciano Huck desistiu de sua candidatura. Para este exemplo, a sequência de eventos fica:

$$E1 \rightarrow E2 \rightarrow E3 \rightarrow E4$$

Os exemplos (68) e (69) são duas manchetes classificadas pelos três anotadores como explícita. Nestes exemplos, mesmo que seja necessário retomar acontecimentos verdadeiros para compreender, a construção da sentença causa uma estranheza no leitor. Uma explicação para isso é que existe uma articulação de dois eventos na sentença, geralmente um satírico e pelo menos um verdadeiro e, quando esses eventos se conectam, o segundo acontecimento articulado com o primeiro evento, gera o absurdo da notícia satírica.

(68) PSDB termina com o governo mas vão continuar transando quando saírem juntos.
[ex.s]

(69) Bolsonaro separa capim para dar a eleitores de Lula mas come antes. [ex.s]

Em (68), (E1) ‘*PSDB termina com o governo*’ é referente ao rompimento político entre o partido político PSDB com o governo (na época, o governo Temer, do MDB). No entanto, quando vinculada com a outra oração (E2) ‘*vão continuar transando quando saírem juntos*’, percebe-se imediatamente que se trata de uma notícia satírica por não ser um referente a um evento verdadeiro. Desse modo, não existe uma sequência de eventos para resgatar os eventos verdadeiros presentes no texto, mas uma oposição de eventos:

E1 *versus* E2

No exemplo (69), o funcionamento é parecido, pois a oração (E1) ‘*Bolsonaro separa capim para dar a eleitores de Lula*’ remete a um evento verdadeiro, quando em 2018, o então candidato à presidente da República, Jair Bolsonaro, decidiu atacar os eleitores do Partido dos Trabalhadores e gravou um vídeo oferecendo capim a eles. Já a segunda oração (E2) ‘*come antes*’, além de se articular com a oração anterior, também faz referência aos (E3) eleitores de Bolsonaro, chamados de ‘gado’ nas redes sociais. Aqui, além da oposição, também existe uma sequência de dois eventos:

E1 *versus* E2 → E3

No entanto, para entender melhor a compreensão da sátira pelo humano nas manchetes analisadas nesta etapa, foi solicitado que cada anotador deveria marcar em qual ponto da sentença acreditava estar o efeito satírico (de humor, ironia, etc.). Assim, como mostra a Tabela 13, das 150 manchetes analisadas, os três anotadores marcaram exatamente o mesmo trecho da sentença em 8 manchetes. Já 33 manchetes foram marcadas no mesmo trecho por dois anotadores e 99 manchetes tinham um trecho marcado em comum. Por fim, em 11 manchetes não houve nenhum trecho em comum entre os anotadores.

Tabela 13 – Concordância da compreensão da sátira entre anotadores

Concordância	Manchetes anotadas
Total concordância	8
Concordância entre dois anotadores	33
Concordância parcial	99
Discordância total	11

Fonte: Elaborada pela autora.

O Quadro 10 exemplifica estas marcações. A sentença (70) é referente à total concordância entre os três anotadores, a sentença (71) é referente à concordância entre dois anotadores, a sentença (72) é referente à concordância parcial e a sentença (73) refere-se à discordância total.

- (70) OAS acelera obra de cela triplex de Lula.
- (71) Após 518 anos, Brasil finalmente se livra da corrupção para sempre.
- (72) Bolsonaro diz que contratou ex e atual mulher porque luta em favor da família
- (73) Padre fará homilia de sete meses para dar tempo de Lula disputar a eleição

Quadro 10 – Exemplos de marcações de sátira nas sentenças

Concordância	Anotador A	Anotador B	Anotador C
Total concordância	cela triplex	cela triplex	cela triplex
Concordância entre dois anotadores	se livra da corrupção para sempre	se livra da corrupção para sempre	–
Concordância parcial	contratou ex e atual mulher porque luta em favor da família	favor da família	porque luta em favor da família
Discordância total	fará homilia de sete meses	disputar a eleição	–

Fonte: Elaborada pela autora.

É interessante observar que apesar da subjetividade ser uma característica presente na compreensão da sátira e, principalmente, considerando a particularidade de cada anotador, como o conhecimento sobre a política brasileira, conhecimento linguístico e conhecimento de mundo, há uma alta concordância de pelo menos um trecho da sentença satírica. Assim, em 93,33% das sentenças há pelo menos um ponto em concordância na delimitação do trecho onde a sátira pode ocorrer conforme o ponto de vista do anotador, sendo que 5,33% das 150 manchetes há uma concordância exata entre os três anotadores, 22% entre dois anotadores e 66% das manchetes há uma concordância de um trecho específico da sentença.



Neste capítulo foram apresentadas as análises realizadas no contexto da descrição das notícias satíricas, tendo como suporte o *subcorpus* de notícias satíricas e suas correspondentes reais. Dessa maneira, toda a elucidação sobre os conceitos de sátira e da abordagem descritiva de conteúdo enganoso levantados nos capítulos anteriores foram essenciais para as análises linguísticas propostas aqui, possibilitando ainda servir como uma fonte inicial de descrição dos fenômenos envolvidos na construção das notícias satíricas.

Ressalta-se ainda, que apesar da análise das notícias ser uma tarefa trabalhosa, tendo em vista a dificuldade de compreensão todos os eventos presentes nas notícias, o olhar humano não é só um fator só complementar, mas é fundamental para a descrição da sátira. Na análise linguística e manual das características presentes nas sentenças das notícias satíricas, foi possível constatar que 60,72% dessas características são semânticas, sendo as mais representativas a quebra de expectativa (QExp) e palavras fora do contexto (FDC). Sobre as características extraídas do *parser* PALAVRAS (BICK, 2000), do dicionário LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001; BALAGE FILHO; PARDO; ALUÍSIO, 2013) e do NILC-Matrix (LEAL, 2021) não apresentaram resultados muito significativos, mas que ainda podem ser vistos como indícios de uma notícia satírica, como a maior incidência

verbal e adverbial, uso de polaridade (positiva, negativa e neutra) e leiturabilidade das notícias.

A importância da interpretação humana ainda pode ser visualizada na anotação realizada, ao se observar que 54,66% das manchetes satíricas anotadas são explícitas, sendo 33,33%, onde 2 dos 3 anotadores concordaram que as manchetes satíricas eram explícitas e 21,33% concordaram de modo unânime que as manchetes satíricas eram explícitas. Considerando a anotação realizada, destaca-se também que 93,33% das sentenças há pelo menos um trecho em concordância, onde a sátira pode ocorrer conforme o ponto de vista e conhecimento de mundo do anotador.

É esse olhar para a língua, para o contexto social e de conhecimento de mundo do leitor que o entendimento é alcançado e é este processo que se espera futuramente que a máquina possa realizar. Vale apontar que outras perspectivas, evidentemente, são possíveis, e outros caminhos poderiam ter sido escolhidos para análise exposta neste capítulo, mas procurou-se evidenciar a análise linguística e a visão do leitor na compreensão da sátira presente nas notícias a partir da anotação realizada. Ainda assim, uma análise linguística e minuciosa e trabalhosa foi realizada neste trabalho, mas uma análise mais profunda e testes em aprendizado de máquina devem ser realizados a partir da tipologia construída e apresentada no capítulo a seguir.

6 TIPOLOGIA DE PISTAS LINGUÍSTICAS EM NOTÍCIAS SATÍRICAS

Neste capítulo é introduzida a tipologia de sinais linguísticos presentes nas notícias satíricas para o português do Brasil. Sua construção ocorre a partir do estudo do mapeamento de características linguísticas de trabalhos já realizados (cf. Capítulo 3), sendo foi possível observar as principais características linguísticas presentes em notícias satíricas em diversos idiomas. Além disso, são consideradas também a análise linguística e os resultados das ferramentas linguístico-computacionais (cf. Capítulo 5). Uma formalização da tipologia aqui descrita, é apresentada no Apêndice B.

6.1 DESCRIÇÃO DA TIPOLOGIA DE SINAIS LINGUÍSTICOS EM NOTÍCIAS SATÍRICAS

A tipologia proposta neste capítulo foi construída a partir das análises linguísticas e dos resultados do processamento automático das notícias – satíricas e reais – que foram realizados no capítulo anterior, além de se basear em trabalhos anteriores que descreveram as notícias satíricas para outras línguas (cf. Capítulo 3). Assim, foram selecionados os principais recursos observados que destacam ou podem colaborar na distinção automática de notícias satíricas e verdadeiras. A tipologia é dividida 8 categorias de características linguísticas gerais que abordam atributos descritivos, morfossintáticos, lexicais, sintático, semânticos, estilísticos, psicolinguísticos e de complexidade textual. Estas categorias são subdivididas em 53 subcategorias de características linguísticas específicas.

6.2 CARACTERÍSTICAS DESCRITIVAS

Como foi visto em alguns trabalhos anteriores (YANG; MUKHERJEE; DRAGUT, 2017; BARBIERI; RONZANO; SAGGION, 2015; ABONIZIO et al., 2020), sabe-se que as *características descritivas* do texto podem ser um recurso para a diferenciação da notícia satírica da notícia real. São consideradas características descritivas do texto, como número de *tokens*, número de *types*, número de caracteres, número de sílabas e número de sentenças, extraídos do NLTK e do spaCy (cf. Seção 4.1.3). A seguir são descritas estas características e quando ela pode ser considerada um fator para identificação de uma possível notícia satírica.

Vê-se que o número de *types* e a média de palavras por sentenças são características com maiores ocorrências nas notícias satíricas quando comparadas às notícias reais:

- Número de types: Quando o número total de palavras diferentes (*types*) for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- Média de palavras por sentenças: Quando a média de palavras por sentenças for **maior** em relação às notícias reais, é provável que a notícia seja satírica.

Porém, o número de tokens, de sentenças, de caracteres, de sílabas e a média de palavras por sentenças tem maiores ocorrências em notícias reais e, conseqüentemente, menor nas notícias satíricas:

- *Número de tokens:* Quando o número total de palavras (*tokens*) for **menor** em relação às notícias reais, é provável que notícia seja satírica.
- *Número de sentenças:* Quando o número total de sentenças for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- *Número de caracteres:* Quando o número total de caracteres for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- *Número de sílabas:* Quando o número total de sílabas for **menor** em relação às notícias reais, é provável que a notícia seja satírica.

Apesar desses indícios, os elementos descritivos da notícia podem não ser um fator significativo da distinção entre notícia real *versus* notícia satírica em ponto de vista linguístico, dado que as principais diferenças estão em níveis mais abstratos da língua, como foi possível observar nas análises anteriores. Porém, vê-se necessária a classificação desses atributos, uma vez que se espera a utilização desta tipologia aqui proposta para fins não só descritivos como também de processamento computacional da sátira e outros recursos figurados da língua.

6.3 CARACTERÍSTICAS MORFOSSINTÁTICAS

As características morfossintáticas são referentes à ocorrência de características presentes na estrutura morfológica e sintática das sentenças das notícias, como as classes gramaticais ou aspectos verbais – tempo, modo, pessoa, etc. Estas características foram extraídas pelo *parser* PALAVRAS (cf. Capítulo 4) e, posteriormente, analisadas quantitativamente conforme a menor e maior ocorrência nos textos satíricos (cf. Capítulo 5). Portanto, quando comparadas as notícias satíricas com as reais, têm-se as seguintes características com maiores ocorrências nas notícias satíricas:

- *Ocorrência de verbos:* Se a incidência verbal, considerando a média do número de **verbos** pelo número total de palavras, for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (74) Depois de **soltar** Paulo Preto, Gilmar **lembrou** a ele de não se **esquecer** do casaquinho porque a noite **estava** fria. [ex.s]
- (75) Temer é terceiro vice a **assumir** Presidência após a redemocratização. [ex.r]

- *Ocorrência de advérbios*: Se a incidência adverbial, considerando a média do número de advérbios pelo número total de palavras, for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (76) **Aparentemente**, lá não é necessário fazer jejum público para executar o próprio trabalho. [ex.s]
- (77) O magistrado *também* concedeu *habeas corpus* de ofício para a filha de Viera, Tatiana de Souza Cremonini. [ex.r]
- *Ocorrência de determinantes*: Se a incidência do uso de determinantes, considerando a média do número de determinantes pelo número total de palavras, for *maior* em relação às notícias reais, é provável que a notícia seja satírica.
- (78) Após **sua** vitória, **o** juiz Sérgio Moro emitiu **um** mandado de prisão preventiva contra **a** petista Gleici. [ex.s]
- (79) Entre petistas, há temor de que seja rejeitado ou um dos ministros faça pedido de vistas, adiando **sua** votação. [ex.r]
- *Incidência do futuro do indicativo*: Se a incidência do tempo verbal futuro do indicativo for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (80) Com o fechamento das 500 agências, as cartas para Lula **deverão** ser entregues apenas após o fim da pena. [ex.s]
- (81) Além da ideia de levarem o protesto contra a prisão arbitrária de Lula para todos os cantos da cidade, com panfletagens, **haverá** uma programação voltada para a formação política. [ex.r]
- *Incidência do futuro do pretérito do indicativo*: Se a incidência do tempo verbal futuro do pretérito do indicativo for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (82) Outra opção **seria** contar o número de vezes que o PT **poderia** ter evitado que tudo isto que está acontecendo no país desde 2014 se não tivesse se corrompido profundamente. [ex.s]
- (83) A informação da decisão, contudo, surpreendeu investigadores, porque Paulo sempre **teria** negado a existência dessas contas. [ex.r]
- *Incidência do presente do subjuntivo*: Se a incidência do tempo verbal presente do subjuntivo for **maior** em relação às notícias reais, é provável que a notícia seja satírica.

- (84) O prefeito se esqueceu do aniversário da cidade e levantou suspeitas de que **esteja** tendo um caso com alguma outra cidade por aí. [ex.s]
- (85) Outra, do ministro Alexandre de Moraes, de manter no STF todos os processos de crimes cometidos por deputados e senadores durante o mandato mesmo que não **tenham** relação com o cargo. [ex.r]
- *Uso de gerúndio*: Se a incidência do tempo verbal do gerúndio no texto for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (86) Já na série “O Mecanismo”, os roteiristas estão **trabalhando** para que, na segunda temporada, o senador Aécio Neves apareça **salvando** um bebê panda na data de 17 de Abril de 2018. [ex.s]
- (87) Lewandowski se disse “estupefato” com a declaração, **acrescentando** que Barbosa não aceitava quem o contrariasse. [ex.r]
- *Uso do infinitivo*: Se a incidência do tempo verbal do infinitivo no texto for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (88) Se o Aécio recebe 2 milhões da JBS e o Janot pede para **prender**, eu que tenho que **resolver**. [ex.s]
- (89) A defesa tem prazo para se **manifestar** sobre o pedido. [ex.s]
- *Uso da 1ª Pessoa do Singular*: Se a incidência de verbos na 1ª Pessoa do Singular no texto for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (90) “**Pensei** que eu já era da família”, desabafou. [ex.s]
- (91) **Fui** sem ganhar hora extra, um imenso sacrifício pessoal. [ex.r]
- *Uso da 3ª Pessoa do Singular*: Se a incidência da 3ª Pessoa do Singular no texto for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (92) Temer **assumi** em uma entrevista que, ao desempregar Dilma, **acabou** pegando gosto pela coisa e **continua** batendo recordes. [ex.s]
- (93) A discussão sobre o *habeas corpus* nem sequer **entrou** no mérito do pedido nesta quinta. [ex.r]
- *Uso da 1ª ou da 3ª Pessoa do Singular*: Se a incidência de verbos na 1ª ou na 3ª Pessoa do Singular no texto for **maior** em relação às notícias reais, é provável que a notícia seja satírica.
- (94) **Havia** boatos de que Paulo Preto assinaria acordo de delação premiada. [ex.s]

- (95) Fui a serviço do Rio de Janeiro, não **era** oficial, mas **era** a serviço do Rio de Janeiro. [ex.r]

Nos casos abaixo, estão listadas as características com maior incidência nas notícias reais se comparadas às notícias satíricas:

- Ocorrência de preposições: Se a incidência do uso de preposições, considerando a média do número de preposições pelo número total de palavras, for *menor* em relação às notícias reais, é provável que a notícia seja satírica.

- (96) Tremor não é problema, pior é o Temer, diz geólogo **sobre** terremoto. [ex.s]

- (97) Bruno Covas é mais conhecido **entre** os homens e os ricos, aponta Datafolha

- Ocorrência de substantivos: Se a incidência do uso de substantivos, considerando a média do número de substantivos pelo número total de palavras, for *menor* em relação às notícias reais, é provável que a notícia seja satírica.

- (98) Joaquim, que sente muitas **dores** na **coluna**, decidiu não ter também na **cabeça**. [ex.s]

- (99) A possível **campanha** poderia ter acabado no **dia** 7 de outubro, **data** de seu **aniversário** de 64 **anos** e também do primeiro **turno** das **eleições** [ex.r]

- Incidência do pretérito imperfeito do indicativo: Se a incidência do tempo verbal pretérito imperfeito do indicativo for **menor** em relação às notícias reais, é provável que a notícia seja satírica.

- (100) Segundo veterinários do Planalto, o cachorro teria se jogado no lago pois **estava** deprimido por ter que conviver com o presidente Michel Temer. [ex.s]

- (101) A expectativa do governo **era** de que a reforma **gerasse** empregos formais e **reduzisse** a informalidade. [ex.r]

- Incidência do pretérito perfeito do indicativo: Se a incidência do tempo verbal pretérito perfeito do indicativo for **menor** em relação às notícias reais, é provável que a notícia seja satírica.

- (102) Temer **assumiu** em uma entrevista que, ao desempregar Dilma, **acabou** pegando gosto pela coisa e continua batendo recordes. [ex.s]

- (103) **Houve** divergência sobre o alcance da medida, mas **prevaleceu** posição de manter no STF somente os processos de crimes cometidos durante o mandato e relacionados ao exercício do cargo. [ex.r]

- Incidência do presente do indicativo: Se a incidência do tempo verbal presente do indicativo for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- (104) Joaquim Barbosa **estuda** abafar algum escândalo para atrair eleitor do PSDB. [ex.s]
- (105) Organizadores e acampados **fazem** balanço positivo, destacando as expressões de apoio. [ex.r]
- Uso do particípio: Se a incidência do tempo verbal do particípio no texto for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- (106) Sergio Moro teria **cantarolado** enquanto esperava o papel sair da impressora. [ex.s]
- (107) Felizmente, esse telefonema, **omitido** pelo delator, foi **recuperado** pela Polícia Federal. [ex.r]
- Uso da 2ª Pessoa do Singular: Se a incidência de verbos na 2ª Pessoa do Singular no texto for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- (108) Cabral e Adriana fazem teologia mas são reprovados no mandamento ‘não **roubarás**’. [ex.s]
- (109) Tu **acha** que, se eu tivesse batido, não tinha uma marquinha, não? [ex.r]
- Uso da 1ª Pessoa do Plural: Se a incidência de verbos na 1ª Pessoa do Plural no texto for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- (110) Lula perdoa PMDB e **estamos** cansados demais para fazer piada. [ex.s]
- (111) Mas nós **vamos** fazer uma campanha franciscana. [ex.r]
- Uso da 3ª Pessoa do Plural: Se a incidência de verbos na 3ª Pessoa do Plural no texto for **menor** em relação às notícias reais, é provável que a notícia seja satírica.
- (112) As falas de Temer em breve **entrarão** para o catálogo da Netflix na categoria *stand up*. [ex.s]
- (113) As escolhas eleitorais de Aécio nos dois últimos pleitos se **demonstraram** equivocadas. [ex.r]

6.4 CARACTERÍSTICAS LINGUÍSTICAS

A compreensão do sentido figurado pode ser alcançada por meio de dispositivos e recursos linguísticos e retóricos agrupados em diferentes categorias. A fim de encontrar esses recursos, realizou-se uma análise linguística no nível da sentença e no nível do documento. Desse modo, abaixo são descritos os aspectos linguísticos identificados na análise apresentada no Capítulo 5. Os objetivos da etapa de análise e identificação de atributos linguísticos das sentenças foram dois: em primeiro lugar, a tarefa de uma análise linguística cuidadosa dos textos das notícias satíricas ajudou a compreender melhor o fenômeno que está sendo descrito; em segundo lugar, pensou-se em como os atributos identificados poderiam ser úteis não só para o mapeamento de recursos linguísticos presentes em notícias satíricas, mas também considerou-se como poderiam ser utilizados como recursos para entrada de treinamento de AM em trabalhos futuros.

A seguir são descritas as 4 características linguísticas (lexicais, sintática, semântica, estilísticas) e as 17 características específicas (cf. Capítulo 5, 7).

Características Lexicais: são referentes ao nível lexical da sentença, divididas em: (i) palavras pejorativas; (ii) vulgarismos; (iii) autoria; (iv) jogo de palavras; (v) *emoticons/emojis* e (vi) repetição.

- *Palavras pejorativas (Pej):* Se houver palavras para ofender ou humilhar alguém, é provável que a notícia seja satírica.

(114) **Coxinhas** e **mortadelas** foram vistos chorando com cara de patos. [ex.s]

- *Vulgarismos (Vul):* Se houver palavras que apresentam vulgaridade e/ou uso de uma linguagem que se opõe às regras formais de uso da língua, é provável que a notícia seja satírica.


(115) A última pesquisa eleitoral mostrou que Lula tem 35%, Bolsonaro tem 15% e o brasileiro tem **merda** na cabeça. [ex.s]

- *Autoria (Aut):* Se houver a assinatura do jornal satírico no texto, é provável que a notícia seja satírica.

(116) A realidade, segundo apurada pelo **Sensacionalista**, é que o empresário foi preso por ter bancado os comerciais do Dollynho. [ex.s]

- *Jogo de palavras (JPa):* Se houver o uso de palavras que, com sons parecidos e significados diferentes que possibilitam muitas interpretações, causando um efeito inesperado e cômico, é provável que a notícia seja satírica.

(117) Gleisi Hoffmann investigada por ouvir disco da **Al Cione**. [ex.s]

- Emoticons/emojis (Emo): Se houver o uso de emojis/emoticons no texto, é provável que a notícia seja satírica.
- (118) Os políticos foram denunciados por Mark Zuckerberg e serão investigados pelo roubo das reações dos emojis na Operação . [ex.s]
- Repetição (Rep): Quando se utiliza mais de uma vez a mesma palavra para enfatizá-la. Esta característica pode ocorrer no nível da sentença e/ou do documento, é provável que a notícia seja satírica.
- (119) O senador **Aécio Neves** está animado para substituir **Aécio Neves** no PSDB. [ex.s]

Característica Sintática: é referente ao nível sintático da sentença, caracterizada pela troca do sujeito.

- Troca do sujeito (TSuj): Se um complemento nominal na notícia verdadeira passa a ser sujeito na notícia satírica, é provável que a notícia seja satírica.
- (120) **Cachorro** de Marcela se jogou no lago por ter que conviver com Temer. [ex.s]
- (121) Marcela Temer pula em lago para salvar seu **cachorro** e afasta segurança que não ajudou. [ex.r]

Características Semânticas: são características linguísticas presentes no nível semântico da sentença. Referem-se sobre os significados não literais do que está sendo dito. São divididas nas categorias: (i) expressões idiomáticas; (ii) palavras fora do contexto de domínio; (iii) ambiguidade; (iv) meme e (v) quebra de expectativa.

- Expressões idiomáticas (EId): Se houver o uso de expressões idiomáticas no texto, é provável que a notícia seja satírica.
- (122) O ex-presidente Lula tem uma **carta na manga** para evitar a prisão. [ex.s]
- Palavras fora do contexto do domínio (FDC): Se houver o uso de palavras que não se encontram dentro de seu contexto semântico, é provável que a notícia seja satírica.
- (123) **Darth Vader** processa **Marina** por dizer que **Bolsonaro** é do **lado negro da Força**. [ex.s]
- Ambiguidade (Amb): Se uma única palavra perfila vários sentidos, entretanto, um dos sentidos é responsável por possibilitar a interpretação engraçada, é provável que a notícia seja satírica.

(124) A polícia do Rio anunciou hoje que não vai punir os policiais envolvidos nas selfies e ainda vai premiar um deles pelo melhor **enquadramento**. [ex.s]

- Meme (Meme): Esta categoria captura características relativas a sinais que revelam o domínio do falante meme, pois pode ser uma terminologia específica utilizada nas mídias sociais. Assim, se houver o uso de meme no texto, é provável que a notícia seja satírica.

(125) Lula **já está** preso na **Austrália**. [ex.s]

- Quebra de expectativa (QExp): Quando há a quebra ou a interrupção do sentido esperado pelo leitor após uma conjunção, ou um conectivo, é provável que a notícia seja satírica.

(126) Huck desiste de candidatura **para** nariz não crescer mais. [ex.s]

Características Estilísticas: são características referentes aos recursos retóricos e estilísticos, conhecidas como figuras de linguagem, presentes na sentença, sendo categorizadas em: (i) eufemismo; (ii) metáfora; (iii) exagero; (iv) personificação e (v) ironia.

- Eufemismo (Euf): Se houver o uso de versões mais suaves de expressões que podem ser consideradas muito grosseiramente em alguns contextos, é provável que a notícia seja satírica.

(127) Animale dá **desconto de até 98% no salário de costureiras**. [ex.s]

- Metáfora (Met): Se houver o uso de dispositivo para designar um objeto ou qualidade mediante uma palavra que designa outro objeto ou qualidade que tem com o primeiro uma relação de semelhança, é provável que a notícia seja satírica.

(128) PSDB homenageou Gilmar Mendes ontem porque **ele é uma mãe para eles**. [ex.s]

- Exagero (Exa): Se houver o uso de palavras e/ou expressões para exagerar algo de tal forma que não pode acontecer na realidade, é provável que a notícia seja satírica.

(129) Prestes a cumprir sua missão, o juiz Sérgio Moro foi visto comprando milho e baralho para alimentar os pombos na praça e jogar buraco com aposentados pelo **resto de sua vida**. [ex.s]

- Personificação (Pers): Se houver o uso de características humanas a seres não humanos ou inanimados, dando um sentido de personificação, é provável que a notícia seja satírica.

(130) Se ele bate com um dedinho em um móvel sem querer, **o móvel é quem fica com dor**. [ex.s]

- *Ironia (Ironia)*: Se houver o uso de palavras e/ou expressões para apresentar o oposto do que realmente deve ser transmitido pelo locutor, é provável que a notícia seja satírica.

(131) Após começar desempregando Dilma, Temer atinge a marca de **13,7 milhões sem emprego**. [ex.s]

6.5 CARACTERÍSTICAS PSICOLINGUÍSTICAS (LIWC)

Assim como abordado por Rubin et al. (2016), Yang, Mukherjee e Dragut (2017) e Salas-Zárate et al. (2017), as características psicolinguísticas das notícias tem o objetivo de explorar quais categorias podem fornecer padrões para a descrição e detecção da sátira. Entre as 23 categorias analisadas, para a detecção de sátira, 19 dessas categorias possuem maiores valores nas notícias satíricas se comparadas às notícias reais, sendo:

- *Alcance (achieve)*: Se a contagem de palavras relacionadas à categoria alcance (por exemplo, *vitória, sucesso, melhor*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (132) Após sua **vitória**, o juiz Sérgio Moro emitiu um mandado de prisão preventiva contra a petista. [ex.s]
- (133) A Associação Nacional dos Delegados da Polícia Federal desejou sorte e **sucesso** ao novo diretor-geral. [ex.r]
- *Afeto (affect)*: Se a contagem de palavras relacionadas à categoria afeto (por exemplo, *feliz, chorou*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (134) “Só queria ficar **abraçado** nele dia e noite, paguei até frete para entrega expressa e o Lula foi liberado sem impostos, mas nada disso adiantou, pois não recebi ele aqui”, explicou. [ex.s]
- (135) **Felizmente**, esse telefonema, omitido pelo delator, foi recuperado pela Polícia Federal. [ex.r]
- *Raiva (anger)*: Se a contagem de palavras relacionadas à categoria raiva (por exemplo, *odiar, matar, irritado*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.

- (136) E nem por isso se **odeiam**, nem por isso passam a ser odiados, nem contestados. [ex.s]
- (137) Tem que ser um que a gente **mata** ele antes dele fazer delação. [ex.r]
- *Ansiedade (anx)*: Se a contagem de palavras relacionadas à categoria ansiedade (por exemplo, *preocupado, com medo*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (138) Rogério 157 está **com medo** de dividir cela com cúpula do PMDB-RJ. [ex.s]
- (139) “Fico muito **incomodada** que as pessoas ficam mais **preocupadas** em tirar foto, com ela já morta, do que respeitar o animal”, disse Lorena Cavalcante Lopes, de 31 anos, advogada. [ex.r]
- *Certeza (certain)*: Se a contagem de palavras relacionadas à categoria certeza (por exemplo, *sempre, nunca*) for maior nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (140) Uma testemunha que não quis se identificar **garantiu** que Sérgio Moro estava muito feliz e mal conseguia se conter. [ex.s]
- (141) Neste caso, em que se vê mais acossado do que **nunca**, tudo que Lula não poderia fazer é permanecer isolado ou fora do jogo. [ex.r]
- *Morte (death)*: Se a contagem de palavras relacionadas à categoria morte (por exemplo, *enterrar, caixão, matar*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (142) Segundo biólogos, a baleia devia estar **morta** há uma semana quando sua carcaça encalhou na praia. [ex.s]
- (143) Quando me **enterrarem**, não vai ser no meu quintal. [ex.r]
- *Discordância (discrep)*: Se a contagem de palavras relacionadas à categoria discordância (por exemplo, *deveria, seria*) for maior nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (144) Outra opção **seria** contar o número de vezes que o PT poderia ter evitado que tudo isto que está acontecendo no país desde 2014 se não tivesse se corrompido profundamente. [ex.s]
- (145) Por quatro votos a um, os magistrados concordaram que ele também **deveria** responder pelo delito de obstrução à Justiça. [ex.r]

- *Sentir (feel)*: Se a contagem de palavras relacionadas à categoria sentir (por exemplo, *sente, toque*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (146) “Só colocar um helicóptero **tocando** aquilo alto em cima da boca que os bandidos saem pedindo arrego”, disse. [ex.s]
- (147) “Mas não **sinto** como uma superação, as coisas foram acontecendo naturalmente comigo”, disse. [ex.r]
- *Amizade (friend)*: Se a contagem de palavras relacionadas à categoria amizade (por exemplo, *amigo, vizinho*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (148) Huck disse que foi convencido por **amigos** e pela família a não concorrer. [ex.s]
- (149) Antes mesmo do julgamento desta quinta-feira, parlamentares já armavam um acordo para que os **colegas** não ficassem presos. [ex.r]
- *Ouvir (hear)*: Se a contagem de palavras relacionadas à categoria ouvir (por exemplo, *ouvir, escutar*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (150) “Imagina passar o dia **ouvindo** o Suplicy cantando *Blowing in the wind*”, disse um amigo. [ex.s]
- (151) Só retornará à rotina quando “alguma autoridade do Conselho Nacional de Justiça, de preferência o corregedor”, **escutar** seu “clamor”.
- *Lar (home)*: Se a contagem de palavras relacionadas à categoria lar (por exemplo, *cozinha, proprietário*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (152) Não é fácil para ele viver na mesma **casa** que Temer. [ex.s]
- (153) Na época, a informação foi confirmada pela própria assessoria do então presidente: “Procurada, a Presidência confirmou que Lula continua proprietário do imóvel”. [ex.r]
- *Inibição (inhib)*: Se a contagem de palavras relacionadas a inibição (por exemplo, *bloquear, restringir*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (154) Sensacionalista pode **encerrar** atividades após Temer não provar que está vivo. [ex.s]

- (155) A Justiça Estadual de São Paulo ainda decidiu **bloquear** o apartamento triplex, que é investigado pela Operação Lava Jato. [ex.r]
- *Intuição (insight)*: Se a contagem de palavras relacionadas à categoria intuição (por exemplo, *pense, saiba*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (156) “Já estou **pensando** em começar a trazer crochê para fazer no Supremo como passatempo”, disse o ministro. [ex.s]
- (157) “Imagino que seja difícil essa ideia de punição”, afirmou ao EL PAÍS o deputado Marcelo Freixo (PSOL). [ex.r]
- *Lazer (leisure)*: Se a contagem de palavras relacionadas à categoria lazer (por exemplo, *cozinhar, conversar, filme*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (158) O Rio, em uma conversa com o Sensacionalista, disse que Crivella dormirá no sofá. [ex.s]
- (159) Reconhecida a validade das gravações feitas de **conversas** nada republicanas com autoridades da República. [ex.r]
- *Religião (relig)*: Se a contagem de palavras relacionadas à categoria religião (por exemplo, *altar, igreja*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (160) **Padre** fará homilia de sete meses para dar tempo de Lula disputar a eleição. [ex.s]
- (161) Ele é **bispo** licenciado da **Igreja** Universal do Reino de Deus, que reprova a festa. [ex.r]
- *Tristeza (sad)*: Se a contagem de palavras relacionadas à categoria tristeza (por exemplo, *chorando, pesar, triste*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (162) Jucá, inclusive, **lamentou** que o grande acordo nacional não ofereça transporte particular. [ex.s]
- (163) O que não significa que eu não fiquei **triste** com tudo o que aconteceu. [ex.r]
- *Palavrões (swear)*: Se a contagem de palavras relacionadas à categoria palavrões (por exemplo, *foda, porra, merda*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (164) Paulo é tão **foda** que Despacito é que fica com Paulo na cabeça. [ex.s]

- (165) “Eu falei deixa de ser um merda, rapaz, e saí de perto”, disse. [ex.r]
- Tentativa (tent): Se a contagem de palavras relacionadas à categoria tentativa (por exemplo, *pode ser, talvez, acho*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (166) Repórter que fazia ‘povo fala’ na rua pergunta a Bruno Covas o que ele **acha** do novo prefeito. [ex.s]
- (167) Por isso, **talvez** eu seja uma das vozes que mais vocalizem o fim do foro privilegiado. [ex.r]
- Polaridade positiva: Se a contagem de palavras relacionadas a palavras positivas (por exemplo, *feliz, bonito, bom*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (168) Por outro lado, alguns brasileiros desempregados afirmam que se sentem mais **felizes** após a reforma. [ex.s]
- (169) A deputada ainda não é coroa, aliás, é muito **bonita**. [ex.r]
- Polaridade negativa: Se a contagem de palavras relacionadas a palavras negativas (por exemplo, *ódio, inútil, inimigo*) for **maior** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (170) “Os tremores em Brasília, por exemplo, não são um problema, **ruim** mesmo lá é o Temer”, disse ao Sensacionalista. [ex.s]
- (171) “É muito **feia**, não faz meu gênero”, disse Bolsonaro.

Em relação às categorias com maiores resultados, apenas 4 das 23 categorias possuem maiores valores nas notícias reais se comparadas às notícias satíricas, sendo:

- Família (family): Se a contagem de palavras relacionadas à categoria família (por exemplo, *filha, pai, tia*) for **menor** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (172) “Mas ele ainda hoje só dorme com a luz acesa e com uma faca debaixo do travesseiro”, diz a **mãe**. [ex.s]
- (173) Michel Temer, acompanhado pela **mulher**, Marcela Temer, busca o **filho** Michelzinho na escola. [ex.r]
- Dinheiro (money): Se a contagem de palavras relacionadas à categoria dinheiro (por exemplo, *auditoria, dinheiro, dever*) for **menor** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.

- (174) As lulas serão pagas por uma iniciativa público/privada em que o público pagará pelas lulas e a iniciativa privada **embolsará** a **grana**. [ex.s]
- (175) O empresário diz que o **dinheiro** era propina. [ex.r]
- *Ver (see)*: Se a contagem de palavras relacionadas à categoria ver (por exemplo, *visão, viu, visto*) for **menor** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (176) Baleia morta abriu os **olhos** na praia de Ipanema ao ouvir preço da água de coco. [ex.s]
- (177) O senador Randolfe Rodrigues, da Rede, **viu** a troca com desconfiança. [ex.r]
- *Trabalho (work)*: Se a contagem de palavras relacionadas a trabalho (por exemplo, *emprego, xerox*) for **menor** nas notícias satíricas em relação às notícias reais, é provável que a notícia seja satírica.
- (178) De acordo com colegas de **trabalho** de Mikael, o jovem já foi flagrado várias vezes jogando *Clash of Clans* no celular durante o **expediente**. [ex.s]
- (179) Bolsonaro **empregou** e promoveu a mulher em gabinete na Câmara. [ex.r]

6.6 CARACTERÍSTICAS DE INTELIGIBILIDADE TEXTUAL (NILC-METRIX)

As características de inteligibilidade textual são as características extraídas do NILC-Matrix e são referentes às métricas de coerência textual que podem ser úteis na identificação de aspectos de inteligibilidade textual presentes em textos de notícias satíricas. Entre as características analisadas, destacam-se:

- *Pronome anafórico do caso reto*: Quando a média de pronome anafórico do caso reto nos pares de sentenças e sentenças adjacentes do texto for menor em relação às notícias reais, é provável que a notícia seja satírica.
- *Pronome demonstrativo anafórico*: Quando Pronome demonstrativo anafórico nos pares de sentenças e sentenças adjacentes do texto for menor em relação às notícias reais, é provável que a notícia seja satírica. A referência anafórica é a relação entre um pronome e o termo anterior que ele substitui.
- *Referentes anafóricos*: Quando a média de candidatos a referente, na sentença anterior ou em até 5 sentenças anteriores (sentenças adjacentes), por pronome anafórico for menor em relação às notícias reais, indicando se a formação de uma cadeia de correferência é facilitada ou não, é provável que a notícia seja satírica.

- Sobreposição de argumentos: Quando a quantidade média de referentes que se repetem nos pares de sentenças e sentenças adjacentes do texto for maior em relação às notícias reais, indicando se a formação de uma cadeia de correferência é facilitada ou não, é provável que a notícia seja satírica.
- Sobreposição de radical de palavras: Se Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Diversidade de preposições: Quando a proporção de preposições distintas em relação ao total de preposições do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Diversidade de pronomes: Quando a proporção de *types* de pronomes em relação à quantidade de *tokens* de pronomes no texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Diversidade de pontuação: Quando a proporção de *types* de sinais de pontuação em relação aos *tokens* de sinais de pontuação do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Diversidade de pronomes relativos: Quando a proporção de *types* de pronomes relativos em relação à quantidade de *tokens* de pronomes relativos no texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Diversidade de verbos: Quando a proporção de *types* de verbos em relação à quantidade de *tokens* de verbos no texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Brunet: Quando a estatística de Brunet for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Honoré: Quando a estatística de Honoré for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Orações subordinadas: Quando a proporção de orações subordinadas em relação a todas as orações do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Preposições por sentença: Quando a média de preposições por sentenças for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Uso de conjunções coordenativas: Quando a proporção de conjunções coordenativas em relação ao total de conjunções do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.

- Uso de conjunções subordinadas: Quando a proporção de conjunções subordinadas em relação ao total de conjunções do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Pronomes em 1ª pessoa: Quando a proporção de pronomes pessoais nas primeiras pessoas em relação a todos os pronomes pessoais do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Pronomes em 3ª pessoa: Quando a proporção de pronomes pessoais nas terceiras pessoas em relação a todos os pronomes pessoais do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Verbos flexionados: Quando a proporção de verbos flexionados em relação a todos os verbos do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Verbos não flexionados: Quando a proporção de verbos não flexionados em relação a todos os verbos do texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Ambiguidade: Quando a proporção entre a quantidade de sentidos que os adjetivos, advérbios, nomes e verbos do texto possuem no TEP (*Thesaurus* Eletrônico do Português) e a quantidade de adjetivos, advérbios, nomes e verbos no texto for maior em relação às notícias reais, é provável que a notícia seja satírica.
- Leiturabilidade: Quando o cálculo do Índice Flesch Brasileiro for considerado fácil, é provável que a notícia seja satírica.

As características de inteligibilidade textual estão relacionadas, mais precisamente, aos aspectos linguísticos presentes nas estruturas linguísticas. Este recurso pode ser uma característica para filtragem no que diz respeito à identificação automática da sátira.



Neste capítulo foi apresentada a tipologia de sinais linguísticos na estrutura das notícias satíricas do português do Brasil. Na verdade, com base nas análises linguísticas propostas no capítulo anterior, que serviram como uma fonte inicial de descrição dos fenômenos envolvidos dessas notícias, foi possível sistematizar fenômenos que ocorrem na tipologia descrita e representada pela Figura 28, que mostra apenas os sinais com maior ocorrência nas notícias satíricas. No entanto, também foi realizada uma formalização com todos os exemplos retratados, contabilizando os sinais com maior ou menor incidência nas notícias satíricas, conforme representado no Apêndice B.

É importante mencionar que os fenômenos encontrados foram mapeados de acordo com aquilo que mais pareceu relevante ou interessante em termos linguísticos. A estratégia de considerar não só os elementos que são mais salientes nas notícias satíricas é importante para também compreender suas diferenças com as notícias reais. Além disso, tem esse mapeamento de ‘indícios’ do que está presente ou não na notícia pode auxiliar a criação de atributos para a detecção automática de notícias falsas e/ou satíricas.

Por fim, apesar do recente interesse pelo desenvolvimento de aplicações e recursos linguísticos desenvolvidos para a detecção de fake news no português do Brasil (GARCIA; AFONSO; PAPA, 2022; CHARLES; RUBACK; OLIVEIRA, 2022; SANTOS; PARDO, 2021), o trabalho sobre notícias paródicas e satíricas ainda é incipiente, lacuna que se acredita começar a preencher por meio desta pesquisa. De fato, espera-se abrir novos caminhos para que se possa oferecer subsídios linguísticos à descrição da sátira para o português do Brasil e que auxilie também ao desenvolvimento de ferramentas para a detecção automática de conteúdo enganoso.

Figura 28 – Tipologia de características linguísticas em notícias satíricas

Características Descritivas	Características Morfosintáticas	Características Lexicais	Característica Sintática	Características Semânticas	Características Estilísticas	Características Psicolinguísticas	Características de Complexidade
Média palavras/sentença	Classes de palavras	Pejorativas	Troca do sujeito	Expressões idiomáticas	Eufemismo	Alcance	Sobreposição de argumentos
Número de types	Advérbios	Vulgarismos		Fora do contexto de domínio	Metáfora	Afeto	Sobreposição de radical de palavras
	Determinantes	Autoria		Ambiguidade	Exagero	Raiva	Diversidade de preposições
	Verbos	Jogo de palavras		Meme	Personificação	Ansiedade	Diversidade de pronomes
	Futuro (indicativo)	Emoicons/emojis		Quebra de expectativa	Ironia	Causa	Diversidade de pontuação
	Futuro do pretérito (indicativo)	Repetição				Morte	Diversidade de pron. relativos
	Presente (subjuntivo)					Discordância	Diversidade de verbos
	Gerúndio					Sentir	Honoré
	Infinitivo					Amizade	Preposições por sentença
	1ª Pessoa do Singular					Ouvir	Média de orações por sentenças
	3ª Pessoa do Singular					Inibição	Conjunções coordenativas
	1ª/3ª Pessoa do Singular					Intuição	Conjunções subordinativas
						Lazer	Verbos flexionados
						Religião	Verbos não flexionados
						Tristeza	Ambiguidade
						Palavrão	
						Tentativa	
						Polaridade positiva	
						Polaridade negativa	

Fonte: Elaborada pela autora.

7 CONCLUSÃO

As notícias satíricas apresentam-se como um tema desafiador. Por um lado, são notícias que divertem seu público e constituem uma dose de humor diário que ajuda a ver o mundo e as notícias do dia-a-dia sob um outro olhar. Por outro lado, a realidade que tem se apresentado num mundo regido pelas redes sociais e por diversas iniciativas de desinformação, e mesmo de notícias reais que alguns anos atrás pareceriam absurdas, traz um desafio ainda maior com a necessidade de identificação dos vários tipos de notícias enganosas.

O quadro teórico utilizado mostra que embora as notícias satíricas sejam um meio de entretenimento e humor, elas podem ser consideradas uma desinformação, desmitificando o conceito de notícias falsas serem apenas aquelas informações criadas intencionalmente para prejudicar alguém.

O trabalho apresentado nesta tese é uma descrição dos principais mecanismos linguísticos subjacentes na construção de notícias satíricas do site mais conhecido desse gênero no Brasil. Buscou-se compreender como acontece a construção de sentido da sátira nestas notícias. Tenta-se assim ajudar na diminuição da reprodução e compartilhamento de conteúdo enganoso em mídias sociais na web. Para tanto, foi criado o SatiriCorpus.Br, um *corpus* composto por notícias satíricas e um *subcorpus* construído a partir de notícias retiradas do *corpus* e notícias verdadeiras proporcionais às satíricas. Este *corpus* é um dos recursos gerados, totalizando 5.048 notícias satíricas categorizadas em 5 categorias e soma 879.091 *tokens* e 36.515 sentenças.

Também foram investigadas pistas linguísticas em notícias satíricas e verdadeiras a partir de uma análise linguística e de ferramentas de PLN, a fim de comparar como elas se constroem. Entre os principais resultados, pôde ser identificado que os aspectos semânticos são as características linguísticas mais utilizadas na construção das notícias satíricas, especificamente a característica de quebra de expectativa e palavras fora do contexto.

Além disso, observou-se que as características morfológicas extraídas do NILC-Matrix e as características psicolinguísticas extraídas do LIWC, embora não apresentem diferenças muito significativas quando comparadas entre notícias satíricas e reais, apontam tendências que podem ser importantes se combinadas.

Diante da análise do *corpus*, assim como a anotação realizada, verificou-se que as principais evidências linguísticas da sátira e de outras formas de linguagem figurada – como a ironia e o sarcasmo – estão presentes no nível pragmático, no conhecimento de mundo do leitor, tornando a tarefa de descrição linguística, sobretudo voltada à aplicação computacional, muito mais complexa. Pois, mesmo que o leitor consiga reconhecer os sinais na constituição da sentença, é necessário acessar certos conhecimentos contextuais e realizar certos cálculos cognitivos e extralinguísticos para que o efeito da notícia seja

alcançado.

Desse modo, ao retomar as duas hipóteses traçadas inicialmente neste trabalho, salienta-se que:

- (H1):** A hipótese inicial foi validada, visto que foi possível identificar diversos fenômenos linguísticos relacionados às características lexicais, sintática, semânticas e estilísticas presentes nas sentenças estudadas no *subcorpus*.
- (H2):** A análise da anotação realizada nas manchetes das notícias satíricas pode validar esta hipótese, indicando que o leitor consegue identificar sinais presentes na estrutura da notícia, mesmo quando precisa retomar conhecimentos extralinguísticos.

Entre as principais contribuições da pesquisa é possível destacar a construção de uma tipologia de características linguísticas de notícias satíricas para o português do Brasil, conforme mostrou o Capítulo 6. Essa tipologia pode colaborar como um recurso linguístico na detecção de notícias satíricas, assim como para conteúdo enganoso. Destaca-se ainda i) a construção SatiriCorpus.Br, um *corpus* com 5.048 notícias satíricas e ii) um *subcorpus* com 150 notícias satíricas e 150 notícias verdadeiras equivalentes às satíricas. Esses recursos linguísticos – tanto a tipologia, quanto o *corpus* e o *subcorpus* – também poderão ser utilizados por demais pesquisas, conforme os objetivos de outras tarefas e estão disponíveis para consulta e acesso à comunidade.

Ainda, foi possível compreender a importância não só de identificar um método linguístico para a análise computacional das notícias satíricas, como também pode ser destacada a relevância do comportamento do usuário das mídias sociais em relação às notícias satíricas, uma vez que a disseminação ou não desse conteúdo como verdade está relacionada em como ele entenderá aquela notícia.

De fato, em Wick-Pedro et al. (2020), observou-se que a maioria dos usuários se utilizam das notícias satíricas para comentar não só sobre a notícia, mas também sobre todo o contexto em que a notícia está inserida. Há ainda uma grande interação dos usuários que compreendem o efeito satírico e de humor presentes nas notícias satíricas. Desse modo, a descrição e identificação de sinais pode ajudar usuários confusos sobre o conteúdo satírico veiculado nas redes sociais.

Aponta-se, ainda, a necessidade do alinhamento de outros recursos linguísticos, como léxicos e bases de dados, com a tipologia trazida nesta tese para a realização de próximos trabalhos em processamento automático de notícias satíricas como em aprendizado de máquina. Assim, ao testar as características lexicais, sugere-se nas características lexicais um léxico de palavrões ou expressões pejorativas e vulgar para a identificação de *Pej* e *Vul* e Um léxico de *emoticons/emojis* para a identificação de *Emo*. Para as características semânticas é importante o auxílio de uma lista de expressões idiomáticas do português do Brasil para a identificação de *EId*, o uso da WordNet.Br para comparar os usos de

uma palavra e comparar seus sentidos para identificar a ambiguidade no texto para o reconhecimento de *Amb*, uso de um léxico de memes para o português do Brasil para *Meme*. Nas características estilísticas, para a identificação de *Ironia* pode ser realizada por meio do cálculo de polaridade das palavras presentes na sentença.

Pensando nos desafios encontrados no desenvolvimento da pesquisa, pode-se citar que compreensão de um conteúdo satírico está intrinsecamente ligada a dispositivos extralinguísticos e ao conhecimento de mundo do leitor, presentes em níveis mais abstratos da língua e, conseqüentemente, de mais complexidade de formalização. Assim, uma das limitações reconhecidas foi justamente observar que para que a análise linguística (cf. Capítulo 5) pudesse ser realizada meticulosamente, uma amostra de texto muito grande dificultaria esse estudo, tornando-a muito mais exaustiva e comprometendo os resultados.

Entendeu-se ainda que dispositivos linguísticos presentes na estrutura do texto das notícias satíricas podem ser um indicativo de sátira. Entretanto, por conter essa característica pragmática, pode existir um limite para a i) a formalização e ii) reconhecimento desses sinais aqui descritos. Assim, como próxima atividade a ser executada é identificar qual é o limite computacional para a análise automática das notícias satíricas.

Um aspecto que não foi abordado nesta tese e que, como trabalho futuro, se mostra promissor é a utilização de Reconhecimento de Entidades Nomeadas (REN). De fato, o que se pôde notar é que as notícias satíricas têm, em sua maioria, pelo menos uma entidade nomeada em seu título. A única categoria que escapa a essa característica é a das notícias satíricas do domínio do comportamento, como em:

- (180) Cresce o número de pessoas com torcicolo após tentar ler conversas no celular do parceiro. [ex.s]
- (181) Indeciso, homem termina de montar prato do almoço na hora do jantar em buffet. [ex.s]

Esse aspecto pode complementar, entre outros, características como repetição (REP) (cf. exemplos (34) e (35)) ou fora de contexto (FDC), como no trabalho de Carvalho et al. (2020). Por exemplo: ‘*Aécio Neves*’, ‘*Joaquim Barbosa*’, ‘*Marina Silva*’ e ‘*Lula*’, por exemplo, são evidentemente entidades nomeadas da política brasileira, enquanto ‘*Ana Furtado*’ é uma entidade nomeada referente ao domínio do entretenimento e ‘*Darth Vader*’ é um personagem fictício. Já ‘*Luciano Huck*’, mesmo sendo um personagem do domínio do entretenimento, aparece em um bom número notícias satíricas analisadas no *subcorpus* inserido no domínio político. De fato, além desse personagem pertencer ao domínio do entretenimento, visto que é um conhecido apresentador brasileiro de programa de auditório, faz parte também do domínio da política – por cogitar se candidatar à presidência algumas vezes. Além disso, é importante ressaltar que é necessário considerar que a representação dessas entidades nomeadas adquirem no imaginário ou no background do leitor. Logo, um falante de outra língua, mas que domine o português do Brasil, ou até mesmo um

falando do português europeu, não terá repertório suficiente para reconhecer determinados personagens da história satírica, embora identifiquem que se tratam de pessoas. Assim, uma ocorrência como a do exemplo (45) poderia ser identificada como sátira por trazer entidades de diferentes domínios, assim como os exemplos (53) e (123). Acredita-se que tal trabalho exigiria a construção de um recurso de REN que incluísse informações sobre domínio a que pertencem essas entidades.

Considerando a lacuna de trabalhos na literatura sobre análise computacional do conteúdo satírico, sobretudo para o português do Brasil, foram construídos neste estudo novos caminhos para estudos da sátira e das notícias satíricas, tanto para linguística, quanto para o PLN. Além disso, como trabalhos futuros, pretende-se investigar e comparar as construções linguísticas de notícias satíricas (182), notícias reais (183) e notícias falsas (184), principalmente de eventos relacionados à pandemia da COVID-19 e das eleições presidenciais no Brasil no período de outubro de 2022.

(182) OMS recomenda tratamento precoce no país: remover Bolsonaro¹

(183) A politização do tratamento precoce que mata a ciência²

(184) Novo estudo sugere que a ivermectina reduz as infecções por Covid em cerca de 75%. Mais de 30 estudos em todo o mundo descobriram que a droga causou melhorias no tratamento!³

Além disso, é interessante para a Linguística e também para o PLN, um aperfeiçoamento do SatiriCorpus.Br com notícias satíricas mais recentes do portal do Sensacionalista, de outros portais satíricos e também de outros tipos de textos que trabalham com a sátira para o português do Brasil, como perfis do Twitter. Por fim, espera-se ainda testar as características aqui levantadas e presentes na tipologia em algoritmos de aprendizado de máquina, assim como investigar automaticamente novos padrões linguísticos presentes em textos satíricos.

¹ Disponível em: <<https://bit.ly/3eqe5ad>>. Acesso em: 10 de out. 2022.

² Disponível em: <<https://bit.ly/3SV1FGh>>. Acesso em: 10 de out. 2022.

³ Disponível em: <<https://bit.ly/3VoOuz2>>. Acesso em: 10 de out. 2022.

REFERÊNCIAS

- ABAD, Carlos Salas. La primera 'fake news' de la historia. **Hist. Comun. Soc.**, v. 24, n. 2, p. 411–431, nov. 2019. Disponível em: <<https://revistas.ucm.es/index.php/HICS/article/view/66268>>.
- ABONIZIO, Hugo Queiroz et al. Language-independent fake news detection: English, portuguese, and spanish mutual features. **Future Internet**, v. 12, n. 5, p. 87, 2020.
- ALLCOTT, Hunt; GENTZKOW, Matthew. Social Media and Fake News in the 2016 Election. **Journal of Economic Perspectives**, v. 31, n. 2, p. 211–236, 2017.
- ALUÍSIO, Sandra; CUNHA, Andre; SCARTON, Carolina. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2016. p. 109–114. Disponível em: <https://doi.org/10.1007/978-3-319-41552-9_10>.
- ANG, Benjamin; ANWAR, Nur Diyanah; JAYAKUMAR, Shashi. Disinformation & Fake News: Meanings, Present, Future. In: JAYAKUMAR, Shashi; ANG, Benjamin; ANWAR, Nur Diyanah (Ed.). **Disinformation and Fake News**. Singapore: Springer Singapore, 2021. p. 3–20.
- ARTSTEIN, Ron. Inter-annotator agreement. In: _____. **Handbook of Linguistic Annotation**. Dordrecht: Springer Netherlands, 2017. p. 297313. Disponível em: <http://link.springer.com/10.1007/978-94-024-0881-2_11>.
- ATTARDO, Salvatore. **Linguistic Theories of Humor**. [S.l.]: Walter de Gruyter, 1994.
- ATTARDO, Salvatore. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. **Rask**, v. 12, p. 3–20, 2000.
- ATTARDO, Salvatore. **Encyclopedia of humor studies**. Los Angeles: SAGE Reference, 2014. OCLC: 984824479.
- BALAGE FILHO, Pedro P.; PARDO, Thiago Alexandre Salgueiro; ALUÍSIO, Sandra M. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. Fortaleza: [s.n.], 2013. p. 215–219. Disponível em: <<https://aclanthology.org/W13-4829>>.
- BARBIERI, Francesco; RONZANO, Francesco; SAGGION, Horacio. Do we criticise (and laugh) in the same way? automatic detection of multi-lingual satirical news in twitter. In: **Proceedings of the 24th International Conference on Artificial Intelligence**. Buenos Aires: AAI Press, 2015. (IJCAI'15), p. 1215–1221.
- BERGSON, Henri. **O Riso: Ensaio sobre o significado do cômico**. 1ª edição. ed. São Paulo: Edipro, 2018.
- BHATT, Shaily et al. Fake News Detection: Experiments and Approaches Beyond Linguistic Features. In: SHARMA, Neha et al. (Ed.). **Data Management, Analytics and Innovation**. Singapore: Springer Singapore, 2021. v. 71, p. 113–128. Disponível em: <https://link.springer.com/10.1007/978-981-16-2937-2_9>.

BICK, Eckhard. **The Parsing System Palavras: Automatic Grammatical Analysis**. Aarhus Denmark; Oakville, Conn: Aarhus University Press, 2000.

BOARINI, Margareth; FERRARI, Pollyana. A desinformação é o parasita do século XXI. **Organicom**, v. 18, n. 34, p. 37–47, 2021. Disponível em: <<https://www.revistas.usp.br/organicom/article/view/170549>>.

BOUCKAERT, Remco R. et al. Weka-experiences with a java open-source project. **Journal of Machine Learning Research**, v. 11, n. 87, p. 25332541, 2010.

BURFOOT, Clint; BALDWIN, Timothy. Automatic satire detection: Are you having a laugh? In: **Proceedings of the ACL-IJCNLP 2009 Conference Short Papers**. Suntec, Singapore: Association for Computational Linguistics, 2009. p. 161–164.

BURKHARDT, Joanna M. History of Fake News. **Library Technology Reports**, v. 53, n. 8, p. 5–9, 2017. Number: 8. Disponível em: <<https://journals.ala.org/index.php/ltr/article/view/6497>>.

CARROLL, John; MINNEN, Guido; BRISCOE, Ted. Parser evaluation. In: **Treebanks**. Amsterdam: Springer, 2003. p. 299–316.

CARVALHO, Paula et al. Situational irony in farcical news headlines. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2020. p. 65–75. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_7>.

CHARLES, Anderson Cordeiro; RUBACK, Livia; OLIVEIRA, Jonice. Fakepedia corpus: A flexible fake news corpus in portuguese. In: _____. **Computational Processing of the Portuguese Language**. Fortaleza: Springer International Publishing, 2022. v. 13208, p. 3745. Disponível em: <https://link.springer.com/10.1007/978-3-030-98305-5_4>.

CLARK, Herbert H.; GERRIG, Richard J. On the pretense theory of irony. **Journal of Experimental Psychology: General**, v. 113, n. 1, p. 121–126, 1984.

COLEMAN, Meri; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, American Psychological Association, US, v. 60, p. 283284, 1975.

DARNTON, Robert. The true history of fake news. **The New York review of books**, 2017.

DEVLIN, Jacob et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <<https://aclanthology.org/N19-1423>>.

D'ONOFRIO, Salvatore. **Os motivos da sátira romana**. Tese (Doutorado) — Universidade de São Paulo, São Paulo, 1968. Disponível em: <<https://repositorio.usp.br/item/000721396>>.

DUARTE, Lélia Parreira. **Ironia e humor na literatura**. 1ª edição. ed. [S.l.]: Alameda, 2006. 360 p.

- DUBAY, William. The Principles of Readability. **CA**, v. 92627949, p. 631–3309, 2004.
- DYNEL, Marta. Beyond a joke: Types of conversational humour. **Language and Linguistics Compass**, v. 3, n. 5, p. 1284–1299, 2009.
- ERMIDA, Isabel. News satire in the press: Linguistic construction of humour in spoof news articles. In: _____. **Language and Humour in the Media**. Newcastle: Cambridge Scholars Publishing, 2012. p. 185–210.
- FIGUEIRA, Filipo Pires. (des)notícia: a (des)construção de um gênero discursivo. **Letras em Revista**, v. 8, n. 01, 2018.
- FINATTO, Maria José Bocorny. Complexidade textual em artigos científicos: contribuições para o estudo do texto científico em português. **Organon**, v. 25, n. 50, 2011.
- FINATTO, Maria José Bocorny; PARAGUASSU, Liana Braga. **Acessibilidade textual e terminológica**. Uberlândia: EDUFU, 2022. (Série E-Classe: Acessibilidade Textual).
- FINATTO, Maria José Bocorny et al. Vocabulário, complexidade textual e compreensão de leitura em ambientes digitais de ensino: uma investigação inicial com alunos do ensino médio. **Texto Livre: Linguagem e Tecnologia**, v. 9, n. 2, p. 6476, 2016.
- FLESCHE, Rudolf. **How to Write Plain English: A Book for Lawyers and Consumers**. Nova Iorque: Barnes & Noble, 1981. Google-Books-ID: oKtOAgAACAAJ.
- FLOOD, Alison. Fake news is 'very real' word of the year for 2017. **The Guardian**, 2017. Disponível em: <<https://www.theguardian.com/books/2017/nov/02/fake-news-is-very-real-word-of-the-year-for-2017>>.
- FORABOSCO, Giovannantonio. Is the concept of incongruity still a useful construct for the advancement of humor research? **Lodz Papers in Pragmatics**, De Gruyter Mouton, v. 4, n. 1, p. 45–62, 2008.
- FOWLER, Alastair. **Kinds of Literature: An Introduction to the Theory of Genres and Modes**. Oxford: Oxford University Press, 1982.
- GARCÍA-DÍAZ, José Antonio et al. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. **Future Generation Computer Systems**, v. 114, p. 506518, 2021.
- GARCÍA-DÍAZ, José Antonio; CÁNOVAS-GARCÍA, Mar; VALENCIA-GARCÍA, Rafael. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. **Future Generation Computer Systems**, v. 112, p. 641657, 2020.
- GARCÍA-DÍAZ, José Antonio; VALENCIA-GARCÍA, Rafael. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. **Complex Intelligent Systems**, v. 8, n. 2, p. 17231736, 2022.
- GARCIA, Gabriel L.; AFONSO, Luis C. S.; PAPA, João P. Fakerecogna: A new brazilian corpus for fake news detection. In: _____. **Computational Processing of the Portuguese Language**. Fortaleza: Springer International Publishing, 2022. v. 13208, p. 5767. Disponível em: <https://link.springer.com/10.1007/978-3-030-98305-5_6>.

GAVIOLLI, Fabiana Moreira. O uso dos emojis por meio do whatsapp nas relações de trabalho. **Anuário Unesco/Metodista de Comunicação Regional**, v. 20, n. 20, p. 247–260, 2016.

GIBBS, Raymond W. **The Poetics of Mind: Figurative Thought, Language, and Understanding**. First edition. Cambridge: Cambridge University Press, 1994.

GIBBS, Raymond W. Irony in talk among friends. **Metaphor and Symbol**, v. 15, n. 1–2, p. 5–27, 2000.

GIBBS, Raymond W.; COLSTON, Herbert L. **Irony in Language and Thought: A Cognitive Science Reader**. Nova Iorque: Lawrence Erlbaum Associates, 2007.

GRAESSER, Arthur C. et al. Coh-Metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, & Computers**, v. 36, n. 2, p. 193–202, 2004. Disponível em: <<https://doi.org/10.3758/BF03195564>>.

GREENBERG, Jonathan. **The Cambridge Introduction to Satire**. Cambridge: Cambridge University Press, 2019.

GRICE, Paul H. Logic and conversation. In: _____. **Speech Acts**. Nova Iorque: New York: Academic Press, 1975. v. 3, p. 41–58.

GUIBON, Gaël et al. Multilingual fake news detection with satire. In: **CICLing: International Conference on Computational Linguistics and Intelligent Text Processing**. La Rochelle, France: [s.n.], 2019. Disponível em: <<https://halshs.archives-ouvertes.fr/halshs-02391141>>.

GUNNING, Robert. **The Technique of Clear Writing**. [S.l.]: McGraw-Hill, 1952. ISBN 978-7-00-001419-0.

HABERMAS, Juergen. **Communication and the Evolution of Society**. Boston: Beacon Press, 1979.

HANSEN, João Adolfo. **A Sátira e o Engenho**. 2a. ed.. ed. Campinas: Ateliê Editorial, 2004.

HIGHET, Gilbert. **The Anatomy of Satire**. Princeton: Princeton University Press, 1962.

HODGART, Matthew. **Satire**. Londres: World University Library, 1969.

HOLMAN, Clarence Hugh; HARMON, William. **A Handbook to Literature**. Califórnia: Macmillan, 1992.

HORVITZ, Zachary; DO, Nam; LITTMAN, Michael L. Context-driven satirical news generation. In: **Proceedings of the Second Workshop on Figurative Language Processing**. Online: Association for Computational Linguistics, 2020. p. 4050.

HOVY, Eduard; LAVID, Julia. Towards a science of corpus annotation: a new methodological challenge for corpus linguistics. **International journal of translation**, v. 22, n. 1, p. 13–36, 2010.

HUANG, Chin-Lan et al. The development of the chinese linguistic inquiry and word count dictionary. **Chinese Journal of Psychology**, Taiwanese Psychological Assn, 2012.

HUTCHEON, Linda. **Uma Teoria da Paródia**. Lisboa, Portugal: Edições 70, 1985.

IDOETA, Paulo Adamo. Hábitos digitais estão atrofiando nossa habilidade de leitura e compreensão? **BBC News Brasil**, São Paulo, 2019. Disponível em: <<https://www.bbc.com/portuguese/salasocial-47981858>>.

IONESCU, Radu Tudor; CHIFU, Adrian Gabriel. Fresada: A french satire data set for cross-domain satire detection. In: **2021 International Joint Conference on Neural Networks (IJCNN)**. Shenzhen, China: IEEE, 2021. p. 18. Disponível em: <<https://ieeexplore.ieee.org/document/9533661/>>.

JOHNSON, Ann; RIO, Esteban del; KEMMITT, Alicia. Missing the joke: A reception analysis of satirical texts. **Communication, Culture and Critique**, v. 3, n. 3, p. 396–415, 2010.

JOHNSON-LAIRD, Philip Nicholas. **Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness**. Harvard: Harvard University Press, 1983.

JURAFSKY, Daniel; MARTIN, James H. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. **Upper Saddle River, NJ: Prentice Hall**, 2008.

KINCAID, J. et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. **Institute for Simulation and Training**, 1975. Disponível em: <<https://stars.library.ucf.edu/istlibrary/56>>.

KNIGHT, Charles A. **The Literature of Satire**. Cambridge: Cambridge University Press, 2004.

KOCH, Ingedore G. Villaça. **O texto e a construção dos sentidos**. 10^a ed.. ed. São Paulo: Ed. Contexto, 2013.

KREUZ, Roger J. The production and processing of verbal irony. **Metaphor and Symbol**, v. 15, n. 1–2, p. 99–107, 2000.

KREUZ, Roger J.; ROBERTS, Richard M. On satire and parody: The importance of being ironic. **Metaphor and Symbolic Activity**, Routledge, v. 8, n. 2, p. 97–109, 1993.

LAGHI, Elena. **Tra Satira e Retorica Social: il Caso Jenus**. Tese (Doutorado) — Università della Svizzera italiana, 2015.

LAMARRE, Heather; LANDREVILLE, Kristen; BEAM, Michael. The Irony of Satire. **International Journal of Press-politics - INT J PRESS-POLIT**, v. 14, p. 212–231, 2009.

LEAL, Sidney Evaldo. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. Tese (Doutorado) — Universidade de São Paulo, 2021. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-16072021-115303/>>.

LEFFA, Wilson J. **Aspectos da leitura**. 1ª ed.. ed. Porto Alegre: Sagra: DC Luzzatto, 1996.

LEVI, Or et al. Identifying nuances in fake news vs. satire: Using semantic and linguistic cues. In: **Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 31–35. Disponível em: <<https://aclanthology.org/D19-5004>>.

LIU, Zhan et al. Detection of satiric news on social media: Analysis of the phenomenon with a french dataset. In: **2019 28th International Conference on Computer Communication and Networks (ICCCN)**. Valencia, Spain: IEEE, 2019. p. 16. Disponível em: <<https://ieeexplore.ieee.org/document/8847041/>>.

LLERA, José Antonio. Prolegómenos para una teoría de la sátira. **Tropelías: revista de teoría de la literatura y literatura comparada**, n. 9-10, p. 281–293, 1999.

MARTIN, Louis et al. CamemBERT: a tasty French language model. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 7203–7219. Disponível em: <<https://aclanthology.org/2020.acl-main.645>>.

MARTINS, T. B. F. et al. Readability formulas applied to textbooks in brazilian portuguese. 1996. Disponível em: <<https://repositorio.usp.br/item/000906089>>.

MCHARDY, Robert; ADEL, Heike; KLINGER, Roman. Adversarial training for satire detection: Controlling for confounding variables. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 660–665. Disponível em: <<https://aclanthology.org/N19-1069>>.

MCNAMARA, Danielle et al. Coh-Metrix: Capturing Linguistic Features of Cohesion. **Discourse Processes - DISCOURSE PROCESS**, v. 47, 2010.

MONTEIRO, Rafael A. et al. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In: VILLAVICENCIO, Aline et al. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2018. v. 11122, p. 324–334. Disponível em: <http://link.springer.com/10.1007/978-3-319-99722-3_33>.

MORETT, Marina Dias. **A sátira do acontecimento jornalístico pelo humor: pseudojornais**. Trabalho de Conclusão de Curso — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

MUECKE, Douglas Colin. **Ironia e o Irônico**. 1ª edição. ed. São Paulo: Perspectiva, 1995. 136 p.

PENNEBAKER, James et al. The development and psychometric properties of liwc2007. 2007.

PENNEBAKER, James W et al. **The development and psychometric properties of LIWC2015**. Austin, TX, 2015.

PENNEBAKER, James W; FRANCIS, Martha E; BOOTH, Roger J. Linguistic inquiry and word count: Liwc 2001. **Mahway: Lawrence Erlbaum Associates**, v. 71, n. 2001, p. 2001, 2001.

PENNYCOOK, Gordon et al. Shifting attention to accuracy can reduce misinformation online. **Nature**, v. 592, n. 7855, p. 590–595, 2021. Disponível em: <<https://www.nature.com/articles/s41586-021-03344-2>>.

PÉREZ-ROSAS, Verónica et al. Automatic Detection of Fake News. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 3391–3401. Disponível em: <<https://aclanthology.org/C18-1287>>.

PESSOA, Breno. Diário de pernambuco. **Sensacionalista e outros sites usam linguagem jornalística para fazer humor e crítica**, 2016. Disponível em: <<https://www.diariodepernambuco.com.br/noticia/viver/2016/11/sensacionalista-e-outros-sites-usam-linguagem-jornalistica-para-fazer.html>>.

PETROV, Slav; DAS, Dipanjan; MCDONALD, Ryan T. A universal part-of-speech tagset. **ArXiv**, abs/1104.2086, 2012.

PUSTEJOVSKY, James; STUBBS, Amber. **Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications**. Annotated edição. Sebastopol, CA: OReilly Media, 2012. 342 p.

QUANDT, Thorsten et al. Fake News. p. 1–6, 2019.

RESENDE, Nair Rodrigues; SOUZA, Ana Cláudia de. A atividade tradutória e a relevância da leitura: legibilidade e leiturabilidade de textos humorísticos traduzidos. **Revista Gatilho**, v. 13, 2011. Disponível em: <<https://periodicos.ufjf.br/index.php/gatilho/article/view/26986>>.

RITCHIE, David. Frame-shifting in humor and irony. **Metaphor and Symbol**, v. 20, n. 4, p. 275–294, 2005.

ROCHA, Arthur de Oliveira. **Paródia satírica e crítica midiática nas notícias fictícias do site Sensacionalista**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, Natal, 2017.

ROCHA, Rejane Cristina. **Da utopia ao ceticismo: a sátira na**. Tese (Doutorado) — Universidade Estadual Paulista “Júlio de Mesquita Filho”, Araraquara, 2006.

ROGOZ, Ana-Cristina; MIHAELA, Gaman; IONESCU, Radu Tudor. SaRoCo: Detecting satire in a novel Romanian corpus of news articles. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**. Online: Association for Computational Linguistics, 2021. p. 1073–1079. Disponível em: <<https://aclanthology.org/2021.acl-short.136>>.

RUBIN, Victoria et al. Fake news or truth? using satirical cues to detect potentially misleading news. In: **Proceedings of the Second Workshop on Computational Approaches to Deception Detection**. San Diego, California: Association for Computational Linguistics, 2016. p. 7–17.

RUBIN, Victoria L.; CHEN, Yimin; CONROY, Nadia K. Deception detection for news: Three types of fakes. **Proc. Assoc. Info. Sci. Tech.**, v. 52, n. 1, p. 1–4, 2015.

SAADANY, Hadeel; ORASAN, Constantin; MOHAMED, Emad. Fake or real? a study of Arabic satirical fake news. In: **Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDMS)**. Barcelona, Spain (Online): Association for Computational Linguistics, 2020. p. 70–80. Disponível em: <<https://aclanthology.org/2020.rdsm-1.7>>.

ŠALAMOUN, Jiří. Definition of satire and irony: refusing to struggle with proteus. **The satire of Ishmael Reed: from non-standard sexuality to argumentation**, p. 30–33, 2019.

SALAS-ZÁRATE, María del Pilar et al. Automatic detection of satire in twitter: A psycholinguistic-based approach. **Knowledge-Based Systems**, v. 128, p. 20–33, 2017.

SALAS-ZÁRATE, María del Pilar et al. A study on LIWC categories for opinion mining in spanish reviews. **Journal of Information Science**, SAGE Publications, v. 40, n. 6, p. 749–760, 2014. Disponível em: <<https://doi.org/10.1177/0165551514547842>>.

SANTOS, Roney Lira de Sales; PARDO, Thiago Alexandre Salgueiro. Structural characterization and graph-based detection of fake news in portuguese. In: **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**. on-line: SBC, 2021. p. 199208. Disponível em: <<https://sol.sbc.org.br/index.php/stil/article/view/17799>>.

SARKAR, Sohan De; YANG, Fan; MUKHERJEE, Arjun. Attending sentences to detect satirical fake news. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 3371–3380.

SCARTON, Carolina; ALUISIO, Sandra Maria. Coh-matrix-port: a readability assessment tool for texts in brazilian portuguese. In: **Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR**. Porto Alegre: [s.n.], 2010. v. 10, n. 1.

SHABANI, Shaban; SOKHN, Maria. Hybrid machine-crowd approach for fake news detection. In: **2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)**. Philadelphia, PA: IEEE, 2018. p. 299306.

SILVA, Amós Coelho. Satira quidem tota nostra est? **Idioma**, n. 19, p. 42–51, 1997.

SILVA, Mateus Araújo. A ironia de sócrates nos diálogos de platão. **Classica - Revista Brasileira de Estudos Clássicos**, v. 7, p. 229–258, 1994.

SILVEIRA, Karine. **Notícias humorísticas: que textos são estes?** Tese (Doutorado) — Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, 2019.

SIMPSON, Paul. **On the Discourse of Satire**. Amsterdam: John Benjamins Publishing Company, 2003.

SINGH, Raj Kishnor. Humour, irony and satire in literature. v. 3, n. 4, p. 63–72, 2012.

SINTRA, Marta Catarina Dias. **Fake News e a Desinformação**. Tese (Dissertação) — Universidade Nova de Lisboa, Lisboa, 2019.

SMITH, E. A.; SENTER, R. J. Automated readability index. **AMRL-TR. Aerospace Medical Research Laboratories (U.S.)**, p. 114, 1967.

SOUZA, Emanuel Barbosa de. **Estudo sociorretórico do Gênero Notícia Satírica: o caso do Portal Meiuorte**. Dissertação (Mestrado) — Universidade Federal do Piauí, Teresina, 2013.

SPERBER, Dan; WILSON, Deirdre. Irony and the use-mention distinction. 1981.

STINSON, Emmett. Satire. In: _____. **Oxford Research Encyclopedia of Literature**. Oxford: Oxford University Press, 2019.

TANDOC, Edson C.; LIM, Zheng Wei; LING, Richard. Defining Fake News: A typology of scholarly definitions. **Digital Journalism**, v. 6, n. 2, p. 137–153, 2018. Disponível em: <<https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143>>.

TAUSCZIK, Yla R.; PENNEBAKER, James W. The psychological meaning of words: LIWC and computerized text analysis methods. **Journal of Language and Social Psychology**, SAGE Publications, v. 29, n. 1, p. 24–54, 2009. Disponível em: <<https://doi.org/10.1177/0261927x09351676>>.

TAVARES, Maria Alice. O verbo no texto jornalístico: notícia e reportagem. **Working Papers em Linguística**, n. 11, p. 123142, 1997.

TOÇOĞLU, Mansur Alp; ONAN, Aytu. Satire detection in turkish news articles: A machine learning approach. In: _____. **Big Data Innovations and Applications**. Cham: Springer International Publishing, 2019. v. 1054, p. 107117. Disponível em: <http://link.springer.com/10.1007/978-3-030-27355-2_8>.

TREVISAN, Michele Kapp; PRÁ, Eduardo Biscayno de; GOETHEL, Mariana Fagundes. Meme: intertextualidades e apropriações na internet. **Revista Observatório**, v. 2, n. 11, p. 277298, 2016.

ULLMAN, B. L. Satira and satire. **Classical Philology**, University of Chicago Press, v. 8, n. 2, p. 172–194, 1913.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 11461151, 2018.

VOUTILAINEN, Ato. **Part-of-Speech Tagging**. Oxford University Press, 2012. Disponível em: <<https://doi.org/10.1093/oxfordhb/9780199276349.013.0011>>.

WANG, William Yang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. **arXiv:1705.00648 [cs]**, 2017. ArXiv: 1705.00648. Disponível em: <<http://arxiv.org/abs/1705.00648>>.

WARDLE, Claire; DERAKHSHAN, Hossein. **Information Disorder : Toward an interdisciplinary framework for research and policy making Information Disorder Toward an interdisciplinary framework for research and policymaking**. Estrasburgo: [s.n.], 2017.

WARDLE, Claire; DERAKHSHAN, Hossein. Thinking about information disorder: formats of misinformation, disinformation, and mal-information. p. 12, 2018.

WEISGERBER, Jean. Satire and irony as means of communication. **Comparative literature studies**, v. 10, n. 2, p. 157–172, 1973.

WICK-PEDRO, Gabriela. **ComentCorpus: identificação e pistas linguísticas para detecção de ironia no português do Brasil**. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2018.

WICK-PEDRO, Gabriela et al. Linguistic analysis model for monitoring user reaction on satirical news for brazilian portuguese. In: **Lecture Notes in Computer Science**. Springer International Publishing, 2020. p. 313–320. Disponível em: <https://doi.org/10.1007/978-3-030-41505-1_30>.

YANG, Fan; MUKHERJEE, Arjun; DRAGUT, Eduard. Satirical news detection and analysis using attention mechanism and linguistic features. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 1979–1989. Disponível em: <<https://aclanthology.org/D17-1211>>.

YOUNG, Dannagal G. et al. Psychology, political ideology, and humor appreciation: Why is satire so liberal? **Psychology of Popular Media Culture**, v. 8, n. 2, p. 134–147, 2019.

ZHANG, Yigeng et al. Birds of a feather flock together: Satirical news detection via language model differentiation. **ArXiv**, abs/2007.02164, 2020.

Apêndices

APÊNDICE A – DIRETRIZES DE ANOTAÇÃO

A compreensão da sátira, na maioria das vezes, está além dos elementos linguísticos presentes na superfície do texto, o que exige, portanto, um contexto extralinguístico e um conhecimento de mundo do leitor para que a sátira seja bem-sucedida. No entanto, podem existir mecanismos linguísticos presentes na estrutura textual que sinalizem o conteúdo satírico do texto.

Este documento contém as diretrizes de anotação para classificar sentenças de notícias satíricas. Para a realização desta anotação, você receberá uma planilha com manchetes satíricas retiradas do site Sensacionalista e deverá classificá-las em três categorias conforme a sua interpretação: 1) Explícita e 2) Implícita.

Os detalhes de cada categoria são os seguintes:

1. Explícita: A sátira é compreendida por meio da relação de pistas presentes no enunciado. Assim, se a manchete satírica contém elementos linguísticos presentes na estrutura do texto que indiquem o efeito satírico – seja um efeito cômico ou de humor, por exemplo – ela deverá ser interpretada como **sentença satírica explícita**.

Alguns exemplos:

- (1) Temer vai ser contratado pelo **Sensacionalista** a partir de 2019.
- (2) **Alckmin** dá um **drible** na **Lava-Jato** e é **convocado** por **Tite**.

2. Implícita: A sátira é compreendida por meio de pistas em um contexto pragmático adicional ao enunciado. Assim, se a manchete satírica **NÃO** contém elementos linguísticos presentes na estrutura do texto que indique o efeito satírico – seja um efeito cômico ou de humor, por exemplo, sendo necessário que o leitor pode detectar o conteúdo satírico com base no conhecimento de mundo em comum com o tema abordado na notícia satírica. Quando isso ocorrer, ela deverá ser interpretada como **sentença satírica implícita**.

Alguns exemplos:

- (3) Áudio de Joesley já comprovava que Temer é muito vivo.
- (4) Gilmar Mendes já abre a geladeira sem pedir e usa banheiro de porta aberta no Jaburu.

A anotação ocorrerá no formulário a ser enviado para cada anotador (como no exemplo na figura abaixo), que deverá selecionar cada opção conforme a sua interpretação.

	A	B	C	D	E	F
1	Manchetes	Explícita	Implícita	Comentários		
2	Bolsonaro não vai a debates porque adolescente não assiste TV aberta	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
3	PSDB homenageou Gilmar Mendes ontem porque ele é uma mãe para eles	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
4	Bom dia para Lula em Curitiba é mais chato que bom dia em grupo de família do Whatsapp, diz pesquisa	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
5	Lula perdoa PMDB e estamos cansados demais para fazer piada	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
6	Após anos de comerciais do Dollynho, dono da Dolly finalmente é preso	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
7	Suíça emite alerta e pode quebrar se tiver que repatriar dinheiro do PSDB	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
8	Joaquim Barbosa desiste da presidência e mira o comando do Caldeirão do Huck	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
9	Aleij vota para 17 de novembro ser o dia de 'santa Carmen Lúcia'	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
10	Após começar desempregando Dilma, Temer atinge a marca de 13,7 milhões sem emprego	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
11	Cachorro de Marcela se jogou no lago por ter que conviver com Temer	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
12	Foro privilegiado cai e renova a esperança do fim da tomada de três pinos	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
13	Temer fecha 500 agências dos Correios para evitar que Lula receba cartas	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
14	Temer vai ser contratado pelo Sensacionalista a partir de 2019	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
15	Com medo de ser preso, morador da Zona Oeste do Rio começa a levar advogado como par no pagode	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
16	PF pede transferência de Lula e 31% dos brasileiros sugerem Palácio do Planalto	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
17	Repórter que fazia 'povo fala' na rua pergunta a Bruno Covas o que ele acha do novo prefeito	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
18	Petista vence BBB e Moro já manda prender	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
19	Joaquim Barbosa estuda abafar algum escândalo para atrair eleitor do PSDB	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
20	Aécio se junta aos 51 milhões que votaram nele e diz que foi ingênuo	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
21	Gleisi Hoffmann investigada por ouvir disco da Al Cione	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
22	Cirque du Soleil abre seleção no Brasil após exibições de malabarismo em defesa de Waack	<input checked="" type="checkbox"/>	<input type="checkbox"/>			
23	Darth Vader processa Marina por dizer que Bolsonaro é do lado negro da Força	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
24	Após deixar triplex, MTST quer ocupar sítio em Atibaia mas ninguém se voluntariou	<input checked="" type="checkbox"/>	<input type="checkbox"/>			

NOTA:

Quaisquer comentários adicionais podem ser feitos na seção de comentários da planilha.

APÊNDICE B – TIPOLOGIA

Características Gerais	Características Específicas	Nome	Reais	Satíricas
Características descritivas	Número de <i>tokens</i>	<i>tokens</i>	+	-
	Número de <i>types</i>	<i>types</i>	-	+
	Número de sentenças	sent	+	-
	Número de caracteres	carac	+	-
	Número de sílabas	silabas	+	-
	Média de palavras/sentença	palavra_sent	-	+
Características morfossintáticas	Ocorrência de advérbios	adv	-	+
	Ocorrência de determinantes	detDetermi	-	+
	Ocorrência de substantivos	subst	+	-
	Ocorrência de preposição	prep	+	-
	Ocorrência de pontuação	pont	+	-
	Ocorrência de verbos	verbos	-	+
	Pretérito imperfeito (indicativo)	p_imp_ind	+	-
	Pretérito perfeito (indicativo)	p_perf_ind	+	-
	Presente (indicativo)	pres_ind	+	-
	Futuro (indicativo)	fut_ind	-	+
	Futuro do pretérito (indicativo)	fut_pind	-	+
	Pretérito imperfeito (subjuntivo)	p_imp_subj	+	-
	Presente (subjuntivo)	pres_subj	-	+
	Gerúndio	gerúndio	-	+
	Infinitivo	infinitivo	-	+
	Particípio	particípio	+	-
	1ª Pessoa do Singular	1PS	-	+
	2ª Pessoa do Singular	2PS	+	-
	3ª Pessoa do Singular	3PS	-	+
	1ª/3ª Pessoa do Singular	1/3PS	-	+
1ª Pessoa do Plural	1PP	+	-	
3ª Pessoa do Plural	3PP	+	-	
Características lexicais	Palavras pejorativas	pej	-	+
	Vulgarismos	vul	-	+
	Autoria	aut	-	+
	Jogo de palavras/trocadilhos	jpa	-	+
	Emoticons/emojis	emo	-	+
	Repetição	rep	-	+
Características semânticas	Expressões idiomáticas	eid	-	+
	Fora do contexto de domínio	fde	-	+
	Ambiguidade	amb	-	+
	Meme	meme	-	+
	Quebra de expectativa	qexp	-	+
Característica sintática	Troca do sujeito	tsuj	-	+

Continua na próxima página...

continuação da página anterior

Características Gerais	Características Específicas	Nome	Reais	Satíricas
Características estilísticas	Eufemismo	euf	-	+
	Metáfora	met	-	+
	Exagero	exa	-	+
	Personificação	pers	-	+
	Ironia	ironia	-	+
Características psicolinguísticas	Alcance (<i>achieve</i>)	alcance	-	+
	Afeto (<i>affect</i>)	afeto	-	+
	Raiva (<i>anger</i>)	raiva	-	+
	Ansiedade (<i>anx</i>)	ansied	-	+
	Causa (<i>cause</i>)	causa	-	+
	Certeza (<i>certain</i>)	certeza	-	+
	Morte (<i>death</i>)	morte	-	+
	Discordância (<i>discrep</i>)	discord	-	+
	Família (<i>family</i>)	familia	+	-
	Sentir (<i>feel</i>)	sentir	-	+
	Amizade (<i>friend</i>)	amizade	-	+
	Ouvir (<i>hear</i>)	ouvir	-	+
	Inibição (<i>inhib</i>)	inib	-	+
	Intuição (<i>insight</i>)	intuicao	-	+
	Lazer (<i>leisure</i>)	lazer	-	+
	Dinheiro (<i>money</i>)	dinheiro	+	-
	Religião (<i>relig</i>)	relig	-	+
	Tristeza (<i>sad</i>)	triste	-	+
	Ver (<i>see</i>)	ver	+	-
	Palavrões (<i>swear</i>)	palavrao	-	+
Tentativa (<i>tent</i>)	tent	-	+	
Trabalho (<i>work</i>)	lazer	+	-	
Polaridade positiva	posit	-	+	
Polaridade negativa	negat	-	+	
Polaridade neutra	neutra	-	+	
Características de complexidade textual	Leiturabilidade	leiturab	+	-
	Pron. anafórico do caso reto	anaf_reto	+	-
	Pron. demon. anafórico	demon_anaf	+	-
	Referência anafórica	ref_anaf	+	-
	Sobreposição de argumentos	sobre_arg	-	+
	Sobreposição de rad. de palavras	sobre_rad	-	+
	Diversidade de preposições	div_prep	-	+
	Diversidade de pronomes	div_pron	-	+
	Diversidade de pontuação	div_pont	-	+
	Diversidade de pron. relativos	div_pron_rel	-	+
	Diversidade de verbos	sobre_verbo	-	+
	Brunet	brunet	+	-

Continua na próxima página...

continuação da página anterior

Características Gerais	Características Específicas	Nome	Reais	Satíricas
	Honoré	honore	-	+
	Preposições/sentença	prep_sent	-	+
	Média orações/sentenças	oracao, <i>ent</i>	-	+
	Distância de dependência	dist_dep	+	-
	Uso de conjunções coordenativas	conj_coord	-	+
	Uso de conjunções subordinadas	conj_subord	-	+
	Verbos flexionados	verbos_flex	-	+
	Verbos não flexionados	verbos_flex	-	+
	Ambiguidade	amb	-	+