UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
CURSO BACHARELADO INTERDISCIPLINAR EM CIÊNCIA E TECNOLOGIA

IZODÁRIA RUTCHELY DE OLIVEIRA MENEZES

ANÁLISE ESTATÍSTICA DE PALAVRAS E SENTENÇAS EM TEXTOS DE JORNAIS POTIGUARES

CARAÚBAS
2022

IZODÁRIA RUTCHELY DE OLIVEIRA MENEZES

ANÁLISE ESTATÍSTICA DE PALAVRAS E SENTENÇAS DE TEXTOS EM JORNAIS POTIGUARES

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Orientador: Prof. Dr. Cid Ivan da Costa Carvalho

CARAÚBAS

2022

MM543a MENEZES, IZODÁRIA RUTCHELY DE OLIVEIRA .
ANÁLISE ESTATÍSTICA DE PALAVRAS E SENTENÇAS EM TEXTOS DE JORNAIS POTIGUARES / IZODÁRIA RUTCHELY DE OLIVEIRA MENEZES. - 2022.
39 f. : il.

Orientador: CID IVAN DA COSTA CARVALHO.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Ciência e Tecnologia, 2022.

1. Linguística de Corpus. 2. Coleta de Dados. 3. Língua. 4. Linguística Computacional. I. CARVALHO, CID IVAN DA COSTA , orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automáto em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Caraúbas
Bibliotecária: Dalvanira Brito Rodrigues
CRB: 15/700

IZODÁRIA RUTCHELY DE OLIVEIRA MENEZES

ANÁLISE ESTATÍSTICA DE PALAVRAS EM TEXTOS DE JORNAIS POTIGUARES

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência e Tecnologia.

Defendida em: 18 de novembro de 2022.

BANCA EXAMINADORA

_____
Prof. Dr. Cid Ivan da Costa Carvalho (UFERSA)
Presidente

_____
Prof. Dr. Mário Gleisse das Chagas Martins (UFERSA)
Membro Examinador

_____
Profa. Dra. Rosilda Sousa Santos (UFERSA)
Membro Examinador

Dedico esse trabalho aos meus pais**, João de Deus de Menezes Neto** e **Geilma Viana de Oliveira Menezes**, pois é graça ao seu esforço que hoje consigo concluir meu curso. São meu porto seguro.

Dedico também ao meu noivo, **Felipe França Damasceno** por sempre me auxiliar e me apoiar da melhor forma possível.

AGRADECIMENTOS

Nessa longa caminhada, recebi o apoio de muitos amigos e familiares, como também, conheci muitas pessoas que se tornaram essenciais durante essa longa jornada.

Agradeço em primeiro lugar, a Deus, pela saúde e determinação para seguir em frente sem desanimar durante a realização deste trabalho, por me ajudar a ultrapassar todos os obstáculos encontrados e permitir que meus objetivos fossem alcançados.

Agradeço aos meus pais João de Deus de Menezes Neto e Geilma Viana de Oliveira Menezes que sempre me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho. Agradeço também ao meu noivo Felipe França Damasceno por todo o apoio, ajuda e palavras de conforto e incentivos que me manteve forte e focada durante toda a realização desse trabalho. Ao meu padrinho Lucas Edson que esteve sempre ao meu lado me auxiliando, aos meus amigos, em especial a Ana Paula, Kegivaneide Pires, Flavia Tarsylla e Maria Leidivânia, por tornar esse período mais leve. Aos meus avós, Gentil Viana, Antônia Jerônimo, Maria Pessoa e Nazinha, por todo o carinho comigo, todas as visitas realizadas e todas as palavras de experiência. Aos meus tios Genilson, Geilza, Genilma, Jarcilma e Genilda, por sempre me receberem super bem.

Agradeço, em especial, meu orientador Dr. Cid Ivan da Costa Carvalho, por está presente desde o início da minha vida acadêmica na UFERSA, por ter desempenhado excelente função de orientar com dedicação e amizade, pelas correções e ensinamentos, por todos os conselhos, pela ajuda e pela paciência.

Agradeço também, à banca examinadora Dr. Mario Gleisse das Chagas Martins e Dra. Rosilda Sousa Santos pelo tempo disponibilizado na leitura e avaliação do meu trabalho.

Agradeço às pessoas com quem convivi ao longo desses anos de curso, que me incentivaram e que certamente tiveram impacto na minha formação acadêmica.

“Educação não transforma o mundo. Educação muda as pessoas. Pessoas mudam o mundo”

Paulo Freire

RESUMO

A linguística de corpus é o estudo aprofundado de dados coletados, sendo eles dados textuais orais ou escritos que precisaram de uma coleta minuciosa, com a finalidade de contribuir para pesquisa de uma determinada língua ou variedade linguística. Geralmente, a coleta é realizada de forma direta por meios digitais, sendo eles: sites, revistas e blogs, ou de forma indireta por meio de programas de busca. A pesquisa objetiva analisar estatisticamente o número de palavras e sentenças de um corpus de textos em jornais potiguares. A importância deste trabalho se dá pela aplicação nos domínios da linguística, da linguística aplicada e da linguística computacional na notabilidade da programação visual e compilação de corpus, mostrando sua relevância para os estudos que visam desenvolver ferramentas computacionais que consistem em copilar e fazer a contagem das palavras em textos escritos. Para isso, fizemos uma pesquisa bibliográfica sobre os conceitos presentes na linguística de corpus, especialmente, do autor Sardinha (2004); depois, os dados na análise foram coletados em notícias de sites online do estado do Rio Grande do Norte. O corpus consiste em 30 textos resultando em 104.464 palavras distribuídas em 4 jornais online, sendo eles: Jornal de Fato, O Mossoroense, Potiguar Notícias e Tribuna do Norte.

Palavras-chave: Linguística de Corpus, Coleta de Dados, Língua, Linguística Computacional.

ABSTRACT

Corpus linguistics is the in-depth study of collected data, whether oral or written textual data that needed a thorough collection, with the purpose of contributing to the research of a particular language or linguistic variety. Generally, the collection is carried out directly by digital means, namely: websites, magazines and blogs, or indirectly through search engines. The research aims to statistically analyze the number of words and sentences of a corpus of texts in Potiguar newspapers. The importance of this work is given by the application in the domains of linguistics, applied linguistics and computational linguistics in the notability of visual programming and corpus compilation, showing its relevance for studies that aim to develop computational tools that consist of copying and counting the words in written texts. For this, we carried out a bibliographical research on the concepts present in corpus linguistics, especially by the author Sardinha (2004); then, the data in the analysis were collected from online news sites in the state of Rio Grande do Norte. The corpus consists of 30 texts resulting in 104,464 words distributed in 4 online newspapers, namely: Jornal de Fato, O Mossoroense, Potiguar Notícias and Tribuna do Norte.

Keywords: Corpus Linguistics, Data Collection, Language, Computational Linguistics.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Dr	Doutor

SUMÁRIO

1. INTRODUÇÃO

A existência de uma coletânea de dados linguísticos naturais, legíveis por computador é central a Linguística de Corpus atual. Existes algumas coletâneas de textos que não são considerados um corpus, como por exemplo:

- Arquivo: depósito de textos sem organização prévia.
- Biblioteca eletrônica: coleção que segue alguns critérios de seleção.

O corpus é uma parte da biblioteca eletrônica, construído a partir de um desenho explicito, com objetivos específicos.

Subcorpus: uma parte de um corpus, pode ser fixa ou mutável (dinâmica, isto é, flexível durante a análise). (SARDINHA, 2004).

A linguística de Corpus trabalha com um conjunto de dados linguísticos textuais coletados criteriosamente com o propósito de servirem para desenvolvimento de pesquisa de uma língua ou uma variedade linguística. Antes, os *corpora* não eram eletrônicos, eram coletados, mantidos e analisados manualmente. Na atualidade, a linguística de corpus tem uma ligação muito forte com as ferramentas computacionais, elas são geralmente utilizadas para organização e extração de informações no corpus para observação e interpretação de dados. O corpus deve ser constituído de dados autênticos e legíveis pela linguagem computacional.

O desenvolvimento de coleta e análise de corpora tornou-se muito importante para a língua, pois as palavras são analisadas gramaticalmente recebendo sua categoria, sendo possível evidenciar e quantificar regularidades, fazer a compilação dos corpora, desenvolver ferramentas para a análise e descrever a linguagem. Isso com um número muito grande de dados, pois a análise não é feita apenas em poucos arquivos.

Para essa área, a linguagem é um sistema de probabilidades, onde certas palavras são mais frequentes que outras. No léxico, as palavras se diferenciam entre aquelas de mais frequências e as de menos frequências em determinado texto ou conjunto de textos. Desse modo, há palavras com frequências de ocorrências bastante raras, e para que haja compreensão dessas frequências, é preciso incorporar uma grande quantidade de palavras.

É importante destacar que para a análise das frequências se faz necessário a compilação do corpus. Para isso, exige-se as etapas a serem seguidas, que são; o projeto do Corpus, que inclui a seleção dos textos e os cuidados com os requisitos; a compilação em si, manipulação, nomeação dos arquivos de textos, e pedidos de permissão de uso, e a anotação.

A seleção dos textos inicialmente, procede-se à seleção dos textos apropriados e

importantes para a pesquisa. Para esta etapa, a definição do tipo de Corpus que está se compilando é relevante. É considerável tomar decisões a respeito ao tamanho e a composição do corpus, e observar os gêneros aos quais eles pertencem. A compilação e manipulação do Corpus consiste no armazenamento em arquivos predeterminados de todos os textos selecionados. Podem-se buscar textos da Web ou mesmo textos impressos.

A nomeação de arquivos e geração de cabeçalhos se dá início depois que todos os textos forem coletados e convertidos em formato "txt", eles devem receber um nome. Destaca-se que essa nomeação deve seguir determinado padrão de forma a facilitar a recuperação posterior de cada texto.

Com essas observações, destaca-se que o objetivo desta pesquisa é apresentar a analise estatística das palavras e sentenças em um corpus de textos de jornais potiguares. Para isso, foram executados três processos específicos: a coleta e o armazenamento dos dados em jornais potiguares; a observação das ocorrências nos dados coletados; a geração dos dados estatísticos em um programa de linguísticas para a geração de dados linguísticos.

Desse modo, este trabalho está dividido em 4 partes. A primeira parte apresenta o referencial teórico da pesquisa, onde será abordado as bibliografias que apresentam informações sobre a linguística de corpus; coleta, armazenamento e pré-processamento; a organização do corpus; a importância da pesquisa quantitativa em linguística; conceitos linguísticos sobre: a etiquetagem de palavras e sentenças linguísticas; compilação de corpora para descrição e análise; medidas estatísticas: média, mediana, moda e desvio padrão. A segunda parte do trabalho expõe sobre a metodologia utilizada para a coleta de dados, onde a mesma está dividida em três tópicos, sendo eles: os métodos de pesquisa; os dados da pesquisa e o procedimento de análise. A terceira parte mostra a analise e os resultados colhidos; e por fim, as considerações finais.

2. REFERÊNCIAL TEÓRICO

2.1 Linguística de corpus

Sardinha (2004, p. 3) define que "a linguística de corpus se ocupa da coleta e da exploração de corpora, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística."

A existência de corpora iniciou-se mesmo sem a utilização de computadores, tendo em vista que o significado de corpus é corpo, conjunto de documentos (baseado do dicionário Aurélio). Na antiguidade, Alexandre o Grande, definiu o corpus Helenístico. Na

Idade Média, existia a copilação de corpus através de citações da Bíblia (SARDINHA, 2004).

Vale ressaltar que o corpus que deu feição aos corpora que existem hoje foi um corpus não computacional, o SEU (Survery of English Usage), que foi copilado por Randolf Quirke juntamente com sua equipe, no ano de 1959, em Londres. Esse corpus foi planejado para conter um total de 1 milhão de palavras, organizado em fichas de papel assim, sendo usado como referência para outros corpora futuros (SARDINHA, 2004).

Com a invenção do computador, mudanças aconteceram. Nos anos 60, os computadores mainframe passaram a equipar centros de pesquisa universitários e foram sendo aproveitados para a pesquisa em linguagem. Com a popularização dos computadores, foi possível o acesso de mais pesquisadores ao processamento de linguagem natural e, concomitantemente, a sofisticação do equipamento permitiu a consecução de tarefas mais complexas, mais eficientemente, sem falar no aumento da capacidade de armazenamento e na introdução de novas mídias, as quais facilitaram a criação e manutenção de corpora em maior número (SARDINHA, 2000).

Esse autor acrescenta que a linguística de corpus vem ganhando espaço. Esse feito acontece não só nos centros acadêmicos, mas também no âmbito empresarial. O interesse vem crescendo em suas aplicações comerciais de estudos baseados em corpora. Conforme Sardinha (2000) fala, a história da Linguística de Corpus está, portanto, intimamente ligada à disponibilidade de corpora eletrônicos.

2.2 Coleta, armazenamento e pré-processamento

Sardinha (2004) afirma que a internet se tornou um vasto depósito de textos dos mais variados tipos. A maneira mais simples de conseguir textos para um corpus na internet é acessar as páginas desejadas e salvar os arquivos no computador. Porém, esse procedimento não é eficiente para a coleta de grandes quantidades de textos.

Para conseguir resultados melhores na retirada de muitos textos da internet, o pesquisador precisa de uma ferramenta que faça o download dos textos automaticamente e que os salve no computador do pesquisador. Esse tipo de ferramenta é chamado de offline browser (SARDINHA, 2004).

Para Sardinha (2004), é importante ter o conhecimento sobre a formatação dos textos, para que assim, eles sejam processados corretamente pelo software escolhido. Para isso, pode-se olhar na terminação do arquivo, se a terminação for .txt indica um arquivo texto De

qualquer forma, além de ver a terminação dó arquivo, o usuário deve visualizar o arquivo completo, essa verificação consiste em abri-lo em um editor de texto.

Durante o pré-processamento do corpus é preciso uma ferramenta mais prática que permite encontrar as sequências de caracteres com rapidez, para facilitar a checagem do andamento da formatação e para que o pesquisador verifique o resultado das expressões regulares que vem aplicando no corpus (SARDINHA, 2004).

2.3 A organização do corpus

Com os textos já coletados e limpos, é feita a organização dos arquivos em uma estrutura coerente. Alguns *corpora* são organizados em pastas simples, outros em pastas hierarquizadas (subpastas), outros em textos salvos em arquivos separados, outros em arquivos que contêm mais de um texto (SARDINHA, 2004).

Sardinha (2004) afirma que é preciso conhecer algumas recomendações para executar essa organização dos corpora. A primeira é que os textos precisam estar em uma pasta principal em que só existam textos do corpus. A segunda sugestão é que seja criada uma subpasta que indique a versão atual do corpus. A terceira dica é que as subpastas criadas reflitam seu conteúdo, ou seja, tenham nomes indicativos do tipo de texto, assunto, modalidade, registro.

Os cabeçalhos são uma parte do arquivo de cada texto do corpus que contém informações sobre o texto, tais como a origem, a data de coleta, o grupo de pesquisa responsável, o tamanho do texto, sistema de transcrição, autoria, participantes. Essas informações deixam explícitos vários pontos importantes do texto, de tal modo que computador possa processá-las (SARDINHA, 2004).

2.4 A importância da pesquisa quantitativa em linguística

A pesquisa quantitativa é a forma de buscar e coletar dados que possam ser transformados em números para uma análise minuciosa posteriormente. Esse é um dos métodos principais para uso de instrumentos quando se tem o objetivo de conhecer e definir práticas e estratégias para determinado ramo de estudos.

Na linguística, a descrição dos dados sobre algum fenômeno presente na pesquisa possuem um objetivo essencial. Os dados e resultados devem ser mencionados de forma mais exata e significativa possível. A explicação desses dados, geralmente depende da possibilidade dos tipos e relações que o pesquisador espera encontrar na coleta (GRIES, 2019).

Segundo Gries (2019, p. 17),

> Essa situação tem a ver com a maneira como a linguística se desenvolveu nos últimos dez anos, mas felizmente esse quadro está constantemente mudando agora. O número de estudos que utilizam metodologia quantitativa tem crescido (em todas as sub-áreas da linguística); o campo está sofrendo uma mudança de paradigma em direção a metodologias mais empíricas. Apesar de a metodologia quantitativa ser bem comum em outras disciplinas, ela ainda encontra bastante resistência no círculo linguístico.

Geralmente, somente os métodos quantitativos podem fazer uma separação de dados específicos e com mais detalhes. Por exemplo, segundo Shirai e Andersen (1995), para um linguista testar uma hipótese aspectual que afirme que os aspectos imperfectivo e perfectivo estão preferencialmente utilizados nos tempos presente e passado, respetivamente, ele teria que testar todos os verbos em todas as línguas. A pesquisa quantitativa auxilia em todo o processo de coleta e analise de dados coletados para que uma definição final seja feita de forma mais detalhada e rápida.

2.5 Conceitos linguísticos

2.5.1 A etiquetagem de palavras

As etiquetas podem identificar divisões do texto como parágrafos e números de linhas ou informações referentes a autoria e origem dos textos, como autor, data de publicação do documento, título, etc., as quais são normalmente obtidas na fase de coleta e registradas em seções na forma de cabeçalhos ou headers. Além disso, os textos podem receber anotações com informações linguísticas de vários tipos, como a classe gramatical de cada palavra ou a estrutura das sentenças. O processo de identificação dos principais constituintes do texto, tais como sintagmas nominais, sintagmas verbais, etc. é conhecido como parsing e deve ser realizado para inserção da anotação de estrutura sintática (EVANS, 2008; JACKSON; MOULINIER, 2002).

Evans (2008) também destaca que qualquer documento preparado para a análise de corpus é apenas uma representação do documento original. Devido a isso, muita informação contextual pode ser perdida na etapa de preparação do texto para um formato exigido pela ferramenta de análise.

A morfologia é "frequentemente definida como o componente da Gramática que trata da estrutura interna das palavras" (SANDALO, 2001, p.181).

Goldsmith (2010) afirma que esta área envolve o estudo do que uma palavra é nas línguas naturais e defende que, na prática, a morfologia abrange o estudo de quatro aspectos autônomos: a identificação do léxico de uma língua, a morfologia, a morfossintaxe, e a decomposição morfológica, ou estudo da estrutura interna das palavras. Nesta seção, trataremos da decomposição morfológica e focalizaremos a morfossintaxe, por serem estes os aspectos da morfologia relevantes para a anotação e busca automática, tarefas contempladas nesta pesquisa.

A morfossintaxe lida com a relação entre os morfemas encontrados dentro de uma palavra e as outras palavras que a cercam na mesma sentença; é uma responsabilidade partilhada entre as disciplinas de sintaxe e morfologia. A decomposição morfológica se preocupa com a estrutura interna das palavras, e quais as regras para formá-las a partir de pequenas unidades - os morfemas, ou seja, é a área da morfologia que estuda a relação dos morfemas no interior da palavra.

Enquanto a palavra é a unidade máxima da morfologia, o morfema é a sua unidade mínima, sendo os morfemas os menores elementos que carregam significado dentro de uma palavra (SANDALO, 2001). Trost (2003) os conceitua como entidades abstratas que expressam características, relacionadas a conceitos semânticos, que são as raízes ou conceitos abstratos, de natureza gramatical.

Segundo Anderson (1982), a flexão é uma operação morfológica com relevância sintática, uma vez que as categorias flexionais refletem uma profunda relação entre a estrutura das palavras e a estrutura das sentenças. Algumas propriedades das palavras individuais são dependentes da sua posição na estrutura sentencial. Pode-se dizer ainda que "A flexão é requerida pela sintaxe da sentença, isto é, um contexto sintático apropriado leva à expressão das categorias flexionais, o que não acontece com a derivação, isenta do requisito “obrigatoriedade sintática”" (GONÇALVES, 2011, p.12).

A morfologia computacional lida com o processamento de palavras, seja na forma escrita ou falada, e tem uma extensa variedade de aplicações práticas (TROST, 2003).

O processamento computacional das palavras inicia com a tokenização. O Tokenizer, também conhecidos como analisador léxico segmenta uma sentença em cadeias de caracteres com significado, chamadas de tokens. A simples abordagem de considerar um token como qualquer sequência de caracteres separados por um espaço em branco pode ser adequada para algumas aplicações, mas pode levar a imprecisões. É preciso considerar sinais de pontuação, vírgulas e hífens. Além disso, existem línguas que permitem o espaço em branco dentro de uma palavra, como por exemplo em pomme de terre no francês (termo

francês para "batata") (JACKSON; MOULINIER, 2002).

As línguas não segmentadas, como muitas línguas orientais, não colocam espaços entre palavras, escrevendo cada token adjacente a outro. Outros idiomas, como alemão, finlandês ou coreano, mantém a maioria dos espaços em branco, mas permitem a criação dinâmica de palavras compostas, como por exemplo Lebensversicherungsgesellschaft (termo alemão para "empresa de seguro de vida"). Estes compostos podem ser considerados como uma única palavra, mas em uma tarefa de recuperação de documentos, é importante que haja uma segmentação identificando cada "parte" dela (JACKSON; MOULINIER, 2002; MIKHEEV, 2003).

A separação do texto em tokens e a delimitação de sentenças podem ser considerados segmentação do texto do tipo low-level, uma vez que são desenvolvidos nos estágios iniciais do processamento do texto. Outra tarefa pode ser consideradas como segmentação high-level: Segmentação intra-sentencial ou inter-sentencial. A primeira envolve a segmentação de grupos linguísticos, como as entidades nomeadas, e segmentação de grupos verbais e grupos nominais, que também é chamado de chunking sintático. A segunda envolve agrupamento de sentenças e parágrafos dentro de tópicos do discurso, que são chamados text tiles (MIKHEEV, 2003).

O processo de segmentação ou tokenização consiste em dividir o texto em unidades distintas ou tokens, que são as menores unidades de um texto. A segmentação também determina limites de palavras, sentenças e parágrafos. Existem muitas ferramentas computacionais, conhecidas como tokenizers ou analisadores léxicos, que fazem este processo de forma automática (MEGERDOOMIAN, 2003; JACKSON; MOULINIER, 2002).

2.5.2 Sentenças linguísticas

A sintaxe é o domínio da análise linguística responsável pela formação de sentenças; do ponto de vista sintático, palavras são unidades que compõem uma sentença, e são agrupadas de acordo com a função gramatical na mesma.

A sintaxe é a disciplina linguística que estuda como as palavras são combinadas para formar sintagmas e como estas se combinam para formar sentenças. Um sintagma é uma unidade sintática construída hierarquicamente, apesar das sentenças pronunciadas ou escritas na língua sejam realizadas pela fonologia ou pela escrita como uma sequência linear de sons ou letras (MIOTO; SILVA; LOPES, 2013).

Na estrutura sintática das sentenças, há dois aspectos distintos, porém interrelacionados. O primeiro compreende as relações gramaticais, como a função de sujeito e a função objeto em uma sentença, e também engloba relacionamentos como modificador-modificado, possuidor-possuído, etc (VAN VALIN JR, 2001; RAPOSO,1992; CHOMSKY, 1993).

O segundo aspecto da sintaxe está relacionado à organização das unidades constituintes das sentenças. Uma sentença não consiste simplesmente de uma sequência de palavras. As palavras estão organizadas dentro de unidades, que por sua vez, estão inseridas em unidades maiores, chamadas de constituintes. A organização hierárquica das unidades na sentença é chamada de estrutura de constituinte (VAN VALIN JR, 2001; RAPOSO, 1992; CHOMSKY, 1993).

No processamento computacional de textos, deve-se ater ao delimitador de sentença que consiste num componente de software cuja tarefa é detectar os limites das sentenças. A realização deste trabalho com precisão deixa de ser simples visto que existe ambiguidade frequentemente presente nos sinais de pontuação que marcam o fim de uma sentença. Por exemplo, o símbolo de ponto "." pode denotar um ponto decimal, ou uma abreviação, além de marcar o final de uma frase. Da mesma maneira, sentenças começam com letras maiúsculas, mas nem todas as palavras que começam com uma letra maiúscula iniciam uma sentença (JACKSON; MOULINIER, 2002).

Para lidar com a ambiguidade dos sinais de pontuação, delimitadores de sentenças muitas vezes utilizam expressões regulares ou regras de exceção. Outras ferramentas de segmentação utilizam técnicas empíricas, baseadas no aprendizado de máquina, e são treinadas em um corpus segmentado manualmente (JACKSON; MOULINIER, 2002).

2.6 Compilação de corpora para descrição e análise

A Linguística Computacional é a área da ciência linguística que cuida de investigar o tratamento computacional da linguagem e das línguas naturais para diversos fins práticos. De acordo com Vieira e Lima (2001, p.1), "é a área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural." A área circunda diversas áreas de pesquisa da "Linguística Teórica e Aplicada, como a Sintaxe, a Semântica, a Fonética e a Fonologia, a Pragmática, a Análise do Discurso, etc." (OTHERO; MENUZZI, 2005, p.22).

Uma coleção de textos, disponível em formato eletrônico ou não eletrônico, com objetivo de fundamentar com dados à investigação linguística, é considerada o corpus da pesquisa. Segundo Kennedy (1998), a maioria dos corpora está armazenada atualmente em formato eletrônico para exploração pelas ferramentas computacionais, mas este tipo de armazenamento começou a ser usado, mais intensamente, apenas a partir dos anos 60. De acordo com Bennet (2010), um corpus é uma grande coleção de exemplos de ocorrências naturais de uma língua armazenada eletronicamente.

Sardinha (2000, p. 325) define a Linguística de Corpus como a área da Linguística que se ocupa da coleta e exploração de corpora e, "como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador", dialogando com a Linguística Computacional em muitos aspectos.

Na linguística de corpus, afirma-se que a linguagem é padronizada (patterned), ou seja, "existe uma correlação entre os traços linguísticos e os contextos situacionais de uso da linguagem", o que se evidencia por colocações, coligações ou estruturas que se repetem significativamente.

Silveira (2008, p. 29) diz que "Compilar – ou criar – um corpus é projetar e codificar uma coleção de documentos coletados dentro de determinados padrões ou exigências, para a realização de estudos linguísticos e computacionais de aprendizagem de máquina". Vários aspectos envolvem a compilação de um corpus e não é consenso entre investigadores quais os atributos que um corpus deve ter. Mesmo quando concordam na utilização de determinados requisitos, pode haver discordâncias na maneira prática de consegui-los. Características como formato, representatividade, organização, anotações utilizadas e padrões de codificação devem ser estudados, discutidos e planejados previamente de acordo com o propósito de cada corpus (KENNEDY, 1998; SANTOS, 1999).

O empreendimento de compilar um corpus eletrônico envolve vários processos, que compreendem a coleta, a preparação, a segmentação e a anotação dos textos. O processo de coleta intenta obter documentos textuais no formato eletrônico que atendam aos requisitos estabelecidos para a composição do corpus.

O processo de preparação envolve a conversão de formatos dos textos. Mesmo já em formato eletrônico, muitos documentos podem não estar no formato adequado para serem lidos pelas ferramentas disponíveis. Em geral, as ferramentas suportam documentos no formato de texto puro, sem nenhuma formatação (formato TXT) ou algum formato específico de análise de corpora. Se o corpus for constituído de arquivos provenientes da Web, haverá necessidade de conversão do formato HTML (Hiper Text Markup Language)

para o texto puro. Os formatos gerados por esse tipo de ferramenta geralmente são PDF ou DOC e também podem demandar a conversão para texto puro (EVANS, 2008).

Além da compilação dos dados, deve-se fazer alguns conceitos linguísticos importantes para o tratamento linguístico como a classificação das palavras, a formação das palavras e a sintaxe, como veremos a seguir.

2.7 Medidas estatísticas

São consideradas as estimativas de medidas mais utilizadas em pesquisas de análise de dados. São valores destinados a resumir o comportamento de uma variável. Essa estatística auxilia na utilização dos dados.

2.7.1 Média aritmética

A média aritmética, ou simplesmente média, é uma medida que funciona como o ponto de "equilíbrio" de um conjunto de dados, é representada pela letra grega μ (devemos ler "mi"). Se usamos dados amostrais para obter a média, é referida como $\bar{X}$ (lemos "Xis barra"). É a medida de tendência central mais popular e pelas suas propriedades matemáticas é bastante usada na Estatística Inferencial (SALSA; MOREIRA; PEREIRA, 2007).

Conforme Salsa, Moreira e Pereira (2007), na média aritmética existem dois casos a serem considerados no cálculo: quando se trata de dados isolados ou não tabelados e quando os dados estão organizados em uma tabela de frequências.

> No primeiro caso citado, se $x_1, x_2, \ldots, x_n$ representam as n observações de uma amostra da variável X, então, sua média aritmética simples é definida como: $\bar{X}$. Assim, definida como:
>
> $$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
>
> Essa equação pode ser escrita de uma forma mais simplificada usando a letra maiúscula sigma $\sum$, que é a notação usada para realizar um somatório. Ou seja, $\sum_{i=1}^{n} x_1$ representa a soma de todos os valores assumidos pela variável X, desde x_1 até x_n . Isso significa que o primeiro índice de x_1 é 1 e segue até o n, este se encontra na parte superior do $\sum$. Assim, podemos escrever a fórmula da média como:
>
> $$\bar{X} = \frac{\sum_{i=1}^{n} x_1}{n}$$
>
> Já para o segundo caso, para facilitar o cálculo da média, pode ser feita uma tabela onde em uma coluna pode-se registrar o produto de

> cada um dos valores assumidos pela variável X por suas respectivas frequências. Quando somarmos esses produtos, o total representa a soma de todos os valores da distribuição. A média será obtida dividindo-se esse total pelo número de observações da amostra, ou seja, pelo somatório das frequências.

2.7.2 Mediana

A mediana é definida como o valor que ocupa a posição central em um conjunto de dados ordenados. Consequentemente, ela tem a propriedade de dividir um conjunto de observações em duas partes iguais quanto ao número de seus elementos: o número de dados que são menores ou iguais à mediana é o mesmo que o número de dados que são maiores ou iguais a ela (SALSA; MOREIRA; PEREIRA, 2007).

De forma contrária da média, a mediana não possui influência de valores extremos, pois ela é uma forma de medida totalmente relacionada com a posição que se encontra no conjunto de dados ordenados. Isso implica que mesmo que a amostra possua dados consideravelmente pequeno e/ou grande, esse valor não irá influenciar no cálculo da mediana, pois esses valores não irá afetar a ordem.

É necessário conhecer a posição que ela se encontra em relação aos elementos ordenados da amostra, para assim encontrar a mediana de um conjunto qualquer de dados estatísticos.

2.7.3 Moda

Por definição, a moda de um conjunto de dados é o valor que aparece mais vezes, ou seja, é aquele que apresenta a maior frequência observada (SALSA; MOREIRA; PEREIRA, 2007).

Em alguns caso a moda não é única. Em uma série estatística de amostras, pode ocorrer de ter duas ou mais observações com um ressalto similar, ou seja, que tenham uma mesma frequência máxima.

Salsa, Moreira e Pereira (2007) fala sobre as distribuições bimodais (duas modas), trimodais ou multimodais. Esses fenômenos são possíveis de acontecer em uma só amostra e que apresentem exatamente o mesmo número de ocorrências. Isso significa que não há moda, pois nenhum dado se destacou; o conjunto é, então, chamado amodal.

2.7.4 Desvio padrão

Segundo Gries (2019), o desvio padrão de uma amostra que possui inúmeros

elementos é representado pela relação do cálculo da diferença entre cada dado da amostra e a sua média, elevando ao quadrado todas essas variações. Ao terminar essa relação, divide essa soma dos quadrados ela quantidade de elementos da amostra menos um e calcula a raiz quadrada.

3. METODOLOGIA DA PESQUISA

3.1 Métodos

O método utilizado para realizar esse trabalho é indutivo de forma exploratória, ou seja, é um método de abordagem responsável por fazer generalização. As pesquisas em Linguística de corpus são, geralmente, exploratórias, pois o estudo *exploratório*, como diz Santos (2008, p. 49), procura coligir as amostras, contar as ocorrências, surpreender-se com os casos que se deparam ao investigar uma amostra. Além disso, procura correlações, experimenta classificações, identifica conjuntos, ou seja, abre sendas, identifica lugares de interesse (para lá voltar ou para outros lá irem).

Por isso, estas pesquisas proporcionam maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses. Desse modo, aprimoram-se as ideias ou a descoberta de intuições feitas anteriormente sobre o corpus. Seu planejamento é, portanto, bastante flexível, de modo que possibilite a consideração dos mais variados aspectos relativos ao fato estudado. Assim, parte-se de algo particular para uma questão mais ampla, ou seja, um aspecto geral em relação aos textos dos jornais e, em relação aos dados, faremos um tratamento estatística dos principais termos presentes no corpus.

3.2 Dados da pesquisa

Para a realização da coleta de dados foi necessário o levantamento de alguns pontos. Como o principal foco de estudo são os textos de jornais, foi definido que seria usado jornais online do estado do Rio Grande do Norte (RN). Uma outra característica relevante para a coleta, foi que esses jornais precisavam estar acessíveis a compilação dos dados digitais para, assim, fazer o trabalho.

Dentro disso, os dados coletados foram de quatro jornais, sendo eles: Jornal de Fato; O Mossoroense; Potiguar Notícias; e Tribuna do Norte. Todos eles disponibilizaram a compilação dos dados para o formato txt.

Já com os jornais definidos, foi visto que, para uma análise quantitativa seria necessário estabelecer um gênero, as diferentes áreas de atuação desse gênero, como também

o ano de publicação, a quantidade de textos para cada área e a quantidade de palavras no corpus. Então, foi designado estudar o gênero notícia na área da cultura, educação, política, saúde e segurança. Foi estabelecido o ano de 2021 para publicação de cada texto que seria trabalhado, e fixou-se a quantidade de 15 textos para cada área. Essa coleta resultou em um corpus de 300 textos com um total de 104.464 palavras.

3.3 Procedimento de análise

Na coleta de dados, procedemos com a nomeação dos arquivos. A nomeação desses arquivos coletados auxiliam na identificação e na análise, pois, a sua denominação possui informações diretas e individuais de cada texto. Como se pode ver no quadro 1, a primeira sigla da etiqueta de cada texto refere-se ao jornal no qual o texto foi coletado. Veja o quadro 1 que apresenta as quatro siglas trabalhadas para cada jornal.

Quadro 1 - Siglas dos Jornais

Parâmetro	**Significado**
df	De Fato
mo	O Mossoroense
pn	Potiguar Notícias
tn	Tribuna do Norte

Fonte: autoria própria.

A segunda sigla da etiqueta de cada texto refere-se ao gênero daquele texto, como foi definido um único gênero para estudo, todos os textos possuem a segunda sigla igual. A sigla no é a abreviação do gênero de notícia.

A terceira sigla da etiqueta de cada texto refere-se a área de atuação, onde foi definida cinco áreas diferentes, onde a tabela 2 mostra as siglas referente a cada área estudada, como mostra o quadro 2.

Quadro 2 - Siglas das Áreas de Atuação

Parâmetro	**Significado**
cu	Cultura
ed	Educação
po	Política
sa	Saúde
se	Segurança

Fonte: autoria própria.

A quarta sigla da etiqueta de cada texto refere-se à sequência da coleta daquela amostra de texto, ou seja, a posição daquele texto na sua área de atuação definida. A tabela 3 exemplifica como essa informação está distribuída.

Quadro 3 - Posição da Amostra Textual

Parâmetro	**Significado**
001	Primeiro texto coletado
002	Segundo texto coletado
003	Terceiro texto coletado
004	Quarto texto coletado
005	Quinto texto coletado

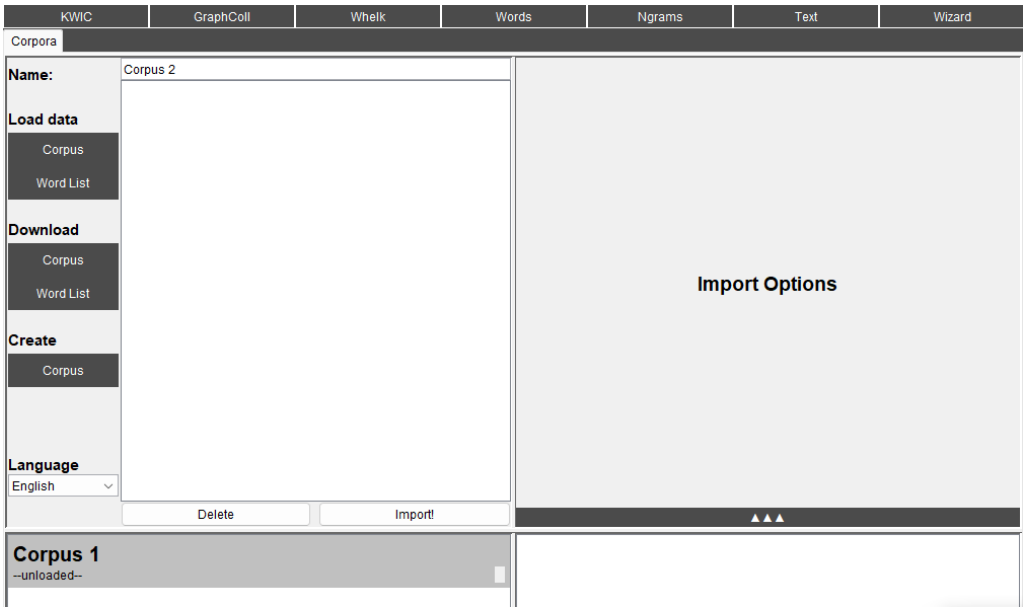
Fonte: autoria própria.

Todas essas informações compõem o título de cada amostra de texto, assim auxiliando na identificação.

Para analisar os dados que foram coletados, foi necessário o uso do programa LancsBox, uma caixa de ferramentas do corpus da Lancaster University, versão 6.0, lançado em 26 de junho de 2021.

Ao abrir o programa LancsBox, a interface do programa apresenta informações como a opção de trabalhar com corpus ou lista de palavras através de carregamento de dados, mostra também a opção de download de um corpus ou lista de palavras, disponibiliza a opção de criação de corpus e a opção de mudança de idioma. Além de oferecer outras configurações onde permite estudo mais elaborado. A figura 1 mostra a interface do programa.

Figura 1 - Interface do Programa LancsBox.



Fonte: Programa LancsBox 6.0.

Inicia-se o processo usando a opção load data, onde será feito o carregamento de dados através da opção corpus, depois mudar o idioma do inglês para o português. Esse processo foi feito separadamente por jornais, ou seja, realizou-se a mesma análise de dados para cada jornal. Após escolher a pasta a ser trabalhada, escolhe-se a opção *import* e os dados serão importados. Na interface, após todas as etapas anteriores, automaticamente é apresentado dados, como a quantidade de files (arquivos), a quantidade de tokens (palavras), a quantidade de types (tipos) e a quantidade de lemas (objeto de definição da palavra).

A primeira análise realizada foi do uso de palavras chave, onde foi gerada uma lista de palavras. No programa, seleciona a opção KWIC e essa ferramenta disponibiliza uma tabela com dados de type, frequência e dispersão.

Foi definido que a análise será realizada nas palavras que apresentem um índice de maior frequência e maior dispersão, para um estudo mais detalhado, foram analisados 10 substantivos com dispersão entre 1,5 a 2,5. Esse processo se repetiu para os 4 jornais.

Após essa análise, foi gerada uma lista de lemas onde foram apresentadas mudanças verbais da mesma palavra. Para cada lema analisado, foi produzida uma etiquetagem. O quadro 4 detalha o significado de cada parte dessa etiqueta.

Quadro 4 – Dados da Etiquetagem das Palavras Analisadas

Parâmetro	**Significado**
N	Inicial da palavra analisada
C	Substantivo Comum
P	Substantivo Próprio
M	Gênero Masculino
F	Gênero Feminino
S	Singular
P	Plural
000	Sequência do substantivo

Fonte: autoria própria

Depois dessa etapa de análise, foi feita a verificação de lemas nas frases onde os substantivos estudados apresentam frequência. Definiu-se que seria feita a escolha de frases diferentes em cada texto e dentro dessas frases foi feita a analise de frequência dos substantivos. Os lemas definem exatamente qual a variação que o substantivo oferece a frase.

4. ANÁLISE E RESULTADO

A coleta de dados foi feita em notícias de jornais potiguares, onde foram escolhidas 5 áreas de atuação dessas notícias em 4 jornais online diferentes. Essa pesquisa totalizou 121.652 tokens coletados.

Após a coleta de dados por jornal, iniciou-se a análise com a etiquetagem adequada para cada jornal, assim, para facilitar na identificação dos textos a serem trabalhados. Então, a primeira etapa da pesquisa é a separação dos substantivos a serem estudados. Foi definido que seria explorado 10 substantivos em cada jornal, sendo esses, os que possuem uma dispersão maior, com escala entre 1,5 a 2,5.

A tabela 1 mostra os dados quantitativos que o Corpus possui.

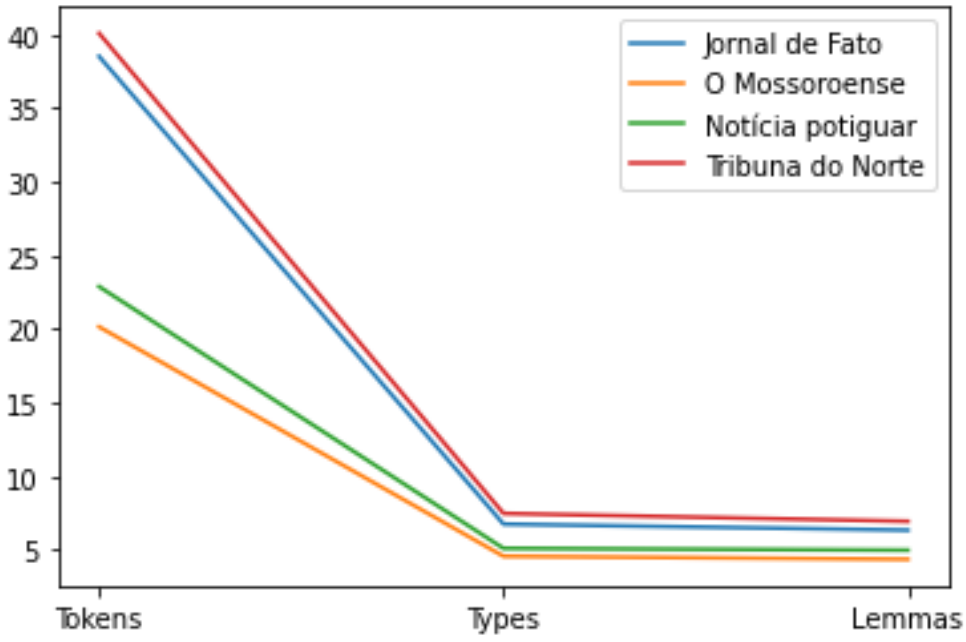
Tabela 1 - Dados quantitativos do Corpus.

Parâmetro	**Jornal de fato**	**O Mossoroense**	**Notícia potiguar**	**Tribuna do Norte**
Tokens	38523	20149	22884	40096
Types	6741	4556	5092	7468
Lemmas	6333	4352	4974	6933

Fonte: autoria própria.

O primeiro jornal a ser analisado foi o Jornal de Fato, logo após foi feita a mesma análise no Jornal o Mossoroense, em seguida repetiu-se o mesmo estudo para o Jornal Notícia Potiguar, e por fim, realizou-se a analise no Jornal Tribuna do Norte.

Gráfico 1 - Curvas quantitativas dos jornais estudados.



Fonte: autoria própria.

O gráfico 1 apresenta as informações que estão tabeladas em forma de curvas. Observa-se que a curva do Jornal de Fato (azul) e a curva do Jornal Tribuna do Norte (vermelha), onde vamos dá enfase nessa pesquisa, mostram comportamentos semelhantes. Esse comportamento implica que os dois jornais possuem uma quantidade próxima de tokens, types e lemmas, mostrando que os dois jornais apresentam textos mais detalhados que, consequentemente, possuem uma quantidade maior de palavras.

Enquanto isso, levando em consideração essa comparação com os jornais que apresentam curvas mais elevadas, o Jornal o Mossoroense (amarela) e o Jornal Notícia Potiguar (verde) apresentam uma curvatura em escala menor, onde mostra que, nessa escala de comparação, os jornais de curvas inferiores possuem uma quantidade menor de características citadas em suas notícias.

Na tabela 2, baseando-se nos conceitos de média, mediana e desvio padrão, encontramos essas características linguísticas nos jornais citados uma relação estatística desses valores.

Tabela 2 - Dados estatísticos nos jornais.

Medidas	Jornal de fato	O Mossoroense	Notícia potiguar	Tribuna do Norte
Média	17.199	9.686	10.983	18.165
Mediana	6.741	4.556	5.092	7.468
Desvio padrão	18.468	9.062087	10.306	18.994

Fonte: autoria própria.

Assim, como vimos na tabela 1 e no gráfico 1, observa-se que a média, a mediana e o desvio padrão nas informações dos Jornal de Fato e Jornal Tribuna do Norte apresentam valores próximos; e os valores para o O Mossoroense e o Jornal Notícia Potiguar também apresentam semelhanças muito próximas. Isso afirma, mais uma vez, que os dois primeiros jornais e dois últimos apresentam quantidade de uso lexicais semelhantes nas suas notícias.

Essas informações sobre os jornais contribuem para vermos a quantidade de substantivos que mais ocorreram no corpus em cada jornal.

O primeiro jornal a ser analisado foi o Jornal de Fato. A tabela 2 mostra os substantivos escolhidos por meio do grau de frequência e dispersão dentro da amostra e com seus respectivos parâmetros da etiquetagem.

Tabela 3 – Substantivos a serem analisados do Jornal de Fato

Substantivo	Frequência	Dispersão	Etiqueta
Saúde	120	1.684181	S-C-F-S-001
Educação	96	2.203595	E-C-F-S-002
Ensino	79	2.749128	E-C-M-S-003
Anos	72	2.203809	A-C-M-P-004
Dia	70	1.613415	D-C-M-S-005
Brasil	69	1.897093	B-P-M-S-006
Governadora	67	1.721626	G-C-F-S-007
Rede	63	1.892575	R-C-F-S-008
Municípios	54	1.882543	M-C-M-P-009
Pessoas	53	2.063534	P-C-F-P-010

Fonte: autoria própria

Usamos como exemplo na amostra a palavra “saúde” que apresentou maior grau de frequência na parcela Jornal de Fato, onde apresentou 120 ocorrências em 41 dos 75 textos coletados. A figura 3 ilustra essas ocorrências dessa palavra no programa Lancsbox, por meio da ferramenta *key words in context* – KWIC, ou seja, palavra chave no contexto.

Figura 2 - Exemplo de ocorrência da palavra saúde.

Fonte: autoria própria

O segundo jornal a ser analisado foi o JORNAL O MOSSOROENSE. A tabela 2 mostra os substantivos a serem analisados, a escolha deles se deu por meio do grau de frequência e dispersão dentro dessa parcela.

Tabela 4 – Substantivos a serem analisados do Jornal O Mossoroense

Substantivo	Frequência	Dispersão	Etiqueta
Saúde	71	1.911266	S-C-F-S-001
Não	66	1.641405	N-C-M-S-002
Municipal	60	1.905723	M-C-M-S-003
Secretaria	43	2.182421	S-C-F-S-004
Vacinação	37	2.416941	V-C-F-S-005
Governo	35	1.974469	G-C-M-S-006
Medidas	34	2.271329	M-C-F-P-007
Trabalho	32	2.302971	T-C-M-S-008
Civil	26	2.429801	C-C-M-S-009
Prefeitura	25	2.277656	P-C-F-S-010

Fonte: autoria própria

Usamos como amostra a palavra “saúde”, por ser o substantivo que apresentou maior grau de ocorrência. Na amostra, o substantivo apresenta 71 ocorrências em 25 dos 75 textos coletados, a figura 3 mostra as ocorrências dessa palavra.

Figura 3 - Ocorrência da palavra “saúde”.

Fonte: autoria própria

Na terceira analise feita, o estudo foi baseado no Jornal Notícia Potiguar. A tabela 2 mostra os substantivos a serem analisados. Os substantivos escolhidos e a etiqueta correspondente a cada um deles está presente na tabela 10. Vale ressaltar que, para a escolha desses substantivos, foi levado em consideração sua frequência e dispersão.

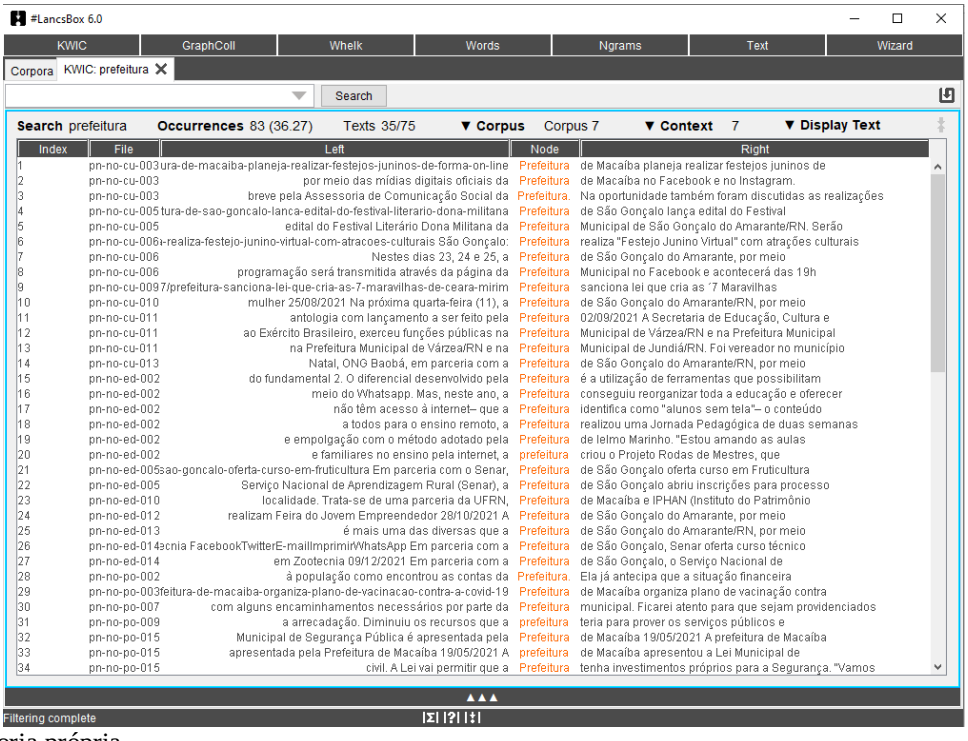
Tabela 5 – Substantivos a serem analisados do Jornal Potiguar Notícias

Substantivo	Frequência	Dispersão	Etiqueta
Prefeitura	83	1.662870	P-C-F-S-001
Anos	77	2.151348	A-C-M-P-002
Municípios	75	2.215685	M-C-M-P-003
Saúde	57	1.773720	S-C-F-S-004
Educação	56	2.307802	E-C-F-S-005
Dia	48	1.797433	D-C-M-S-006
Ensino	42	2.681689	E-C-M-S-007
Cidade	40	1.810146	C-C-F-S-008
Norte	37	2.142274	N-C-M-S-009
Câmara	36	1.969564	C-C-F-S-010

Fonte: autoria própria

Usamos como amostra a palavra PREFEITURA, por apresentar maior ocorrencia. Na amostra, o substantivo apresenta 83 ocorrências em 35 dos 75 textos coletados, a figura 3 mostra as ocorrências dessa palavra.

Figura 4 - Ocorrência da palavra “prefeitura”.



Fonte: autoria própria

A quarta analise estuda o Jornal Tribuna do Norte, onde se realizou o mesmo processo de pesquisa onde os dados de frequência, dispersão e etiqueta estão disponíveis na tabela 12.

Tabela 6 – Substantivos a serem analisados do Jornal Tribuna do Norte

Substantivo	Frequência	Dispersão	Etiqueta
Natal	96	1.597745	N-P-F-S-001
Segurança	80	2.126154	S-C-F-S-002
População	58	1.954907	P-C-F-S-003
Social	49	1.653126	S-C-M-S-004
Todos	47	1.644348	T-C-M-S-005
Forma	41	1.788632	F-C-F-S-006
Trabalho	41	1.615755	T-C-M-S-007
Ação	37	2.055220	A-C-F-S-008
Situação	37	2.123177	S-C-F-S-009
Mesmo	37	1.712252	M-C-M-S-010

Fonte: autoria própria

Foi definida a palavra “Natal” como amostra, por apresentar maior ocorrência na amostra. O substantivo apresenta 96 ocorrências em 38 dos 75 textos coletados, a figura 3 mostra as ocorrências dessa palavra.

Figura 5 - Ocorrência da palavra “Natal”.

Fonte: autoria própria

5. CONSIDERAÇÕES FINAIS

Neste trabalho, podemos considerar que a compreensão e a análise de um corpus por completo é um grande desafio. Este trabalho teve como base o autor Sardinha o qual explica como realizar a coleta de dados para um corpus, onde é visto, de forma minuciosa, como coletar, armazenar e analisar os textos. No entanto, quando analisamos um corpus percebemos que os valores linguísticos das ocorrências apresentam mais do que uma significação única.

Envolvendo de forma totalmente direta a linguística de corpus com a estatística, foi possível ver que a coleta feita nos quatro jornais potiguares proporcionou um corpus online de tamanho médio. Nesse estudo, realizou-se a análise de ocorrências dos substantivos em cada jornal de forma separada. Foi visto que o Jornal de Fato e o Jornal Tribuna do Norte apresentam curvas estatísticas aproximadas, mostrando que tem altos índices de tokens, types e lemmas. Isso implica que os dois jornais trabalham com notícias mais detalhadas sobre os fatos, apresentando, consequentemente, mais número de palavras e mais informações. Comparando-os com o Jornal Notícia Potiguar e O Mossoroense, os quais apresentam as curvas com os níveis mais baixo, significa então que os jornais mostram uma quantidade menor de informação.

Além disso, análise mostra que os substantivos com maior grau de frequência pertencem ao mesmo campo semântico, ou seja, a palavra encontra-se totalmente relacionada ao campo de atuação da área que foi escolhida naquele determinado jornal. No estudo estatístico, a media, mediana e desvio padrão dos mesmos jornais confirmam valores aproximados entre os jornais. Isso comprova que os dois jornais, tanto o Jornal de Fato como o Tribuna do Norte trazem mais detalhes em suas notícias.

Consideramos que essas informações contribuem para compreendermos que a estatística auxilia na análise de corpus linguísticos. Outro fato muito importante é que um corpus como fonte de informação registra a linguagem natural utilizada por falantes e escritores da língua em situações reais. E a investigação da frequência de ocorrência dos substantivos, por exemplo, faz parte do conhecimento linguístico e isso é visto na frequência atestada que se pode estimar a probabilidade teórica.

REFERÊNCIAS

ANDERSON, S. R. Where´s morphology. Linguistic Inquiry, v. 13, 1982.

BARBOSA, Plínio. Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação. Rev. Est. Ling., Belo Horizonte, v. 20, n. 1, p. 11-27, jan./jun., 2012.

BENNET, G. R. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. Michigan: Michigan ELT, 2010.

BATES, M.; Weischedel, R. “Challenges in Natural Language Processing”. Cambridge University
Press, 2006, 312p.

CHOMSKY, N. Lectures on government and binding. The Pisa lectures. 7 ed. Berlim; New Yortk: Mouton de Gruyter, 1993.

EVANS, D. Information about Corpus building and investigation: a on-line information pack about corpus investigation techniques for the Humanities. Birmingham: Centre for Corpus Research/University of Birmingham, 2008. Disponível em:<http://www.birmingham.ac.uk/documents/collegeartslaw/corpus/intro/unit2.pdf >. Acesso em: 04 set. 2022.

SANDALO, M.F. Morfologia. In: MUSSALIM, F. BENTES, A.C. Introdução à linguística. 9 ed. São paulo: Cortez Editora, 2001.

GOLDSMITH, J.A. Segmentation and Morphology. In: CLARK, A.; FOX, C.; LAPPIN, S. (Editores). The Handbook of Computational Linguistics and Natural Language Processing. Willey-BackWell, 2010.

GONÇALVES, C.A. Iniciação aos estudos morfológicos. Flexão e derivação em português. São Paulo: Contexto, 2011.

GRIES, Stefan Th. Estatística com R para a linguística. Tradução: Heliana R. Mello, Crysttian A. Paixão, André L. E. Souza e JúliaZara. Belo Horizonte: FALE/UFMG, 2019.

JACKSON, P.; MOULINIER, I. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

KAY, M. Introduction. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

KENNEDY, G. An introduction to Corpus linguistics. London: Longman, 1998.

McENERY, T. Corpus Linguistics. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

MEGERDOOMIAN, K. Text mining, Corpus building, and testing. In:FARGHALY, Ali Ahmed Sabry (Ed.). Handbook for language engineers. Standford : CSLI, 2003. pp.14.

MIKHEEV, A. Text Segmentation. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003

MIOTO, C.; SILVA, M.C.F.; LOPES, R. Novo Manual de Sintaxe. São Paulo: Contexto, 2013.

OTHERO, G.A. Linguística Computacional: Uma breve introdução. Letras de Hoje, Porto Alegre v.41, n.2, 2006.

OTHERO, G.A.; MENUZZI, S.M. Linguística Computacional: teoria & prática. São Paulo: Parábola Editorial, 2005.

RAPOSO, E.P. Teoria da Gramática à faculdade da Linguagem. Lisboa: Caminho, 1992.

SALSA, Ivone; MOREIRA, Jeanete; PEREIRA, Marcelo. Medidas de tendência central: média, mediana e moda. 27 de Julho de 2007. Apresentação do Power Point. Disponível em: https://d1wqtxts1xzle7.cloudfront.net/32647977/4426477-Matematica-e-Realidade-Aula-08-551-with-cover-page-v2.pdf?Expires=1663807730&Signature=Jgw3rFZzhA0jM9pAark5vGEgbTT-ZaTilKO3M~vhVxuxzE9fd9GmNLo2bUI3ktCOMrCX93V7bsijbuhhY1X7TinXWUIEH4yGn6xzWOKSCUc4OGcfdeFNzpQmYUS4WIHIpEfGADly4Fewsh~3NUz7p3ZEE8VvWrmzQtskA-5DuUOZw8S2ardZbjwgHNxj9DPfavGW7Zg6IhPrnVWKB9NsEs4XsYFeNVP46oFcx6sHEIPjrfYRGGlj3Nv7uYzHl4l8CQdwrSzhd9qT~NJ5zspbFZnSXGiIuW3UMkOEw0kO3~tspk4vtfEqisP31tZHj55d5LdcIqTPUKrAb8YaifA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>. Acesso em: 21 de setembro de 2022.

SANTOS, D. Disponibilização de corpora através da WWW. In Palmira Marrafa & Maria

Antónia Mota (eds.), Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística (Lisboa, 25-27 de Maio de 1998), Lisboa: Colibri, 1999.

SARDINHA, T. B. Linguística de corpus: histórico e problemática. D.E.L.T.A., São Paulo, 2000.

SARDINHA, Tony Berber. Linguística de Corpus. Barueri – São Paulo: Manole, 2004. 410p.

SHIRAI, Y.; ANDERSEN, R. W. The Acquisition of Tense-Aspect Morphology: A Prototype Account. Language, v.71, issue 4 (Dec., 1995), 1995.

SILVEIRA, F.P. Integração de ferramentas para compilação e exploração de corpora. 2008. 101 f. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2008.

TROST, H. Morphology. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

VAN VALIN JR, R.D. An Introduction to Syntax. New York: Cambridge University Press. 2001.

VIEIRA, R.; LIMA, V. L.S. Lingüística Computacional: princípios e aplicações. In: Ana Teresa Martins; Díbio Leandro Borges (Org.). SBC - Jornadas de Atualização em Inteligência Artificial (JAIA). Fortaleza, 2001, v. 3.

[1] Trecho meramente ilustrativo retirado da obra: DUARTE, E. N. D.; SILVA, A. K. A. da; COSTA, S. Q. da. Gestão da informação e do conhecimento: práticas de empresa “excelente em gestão empresarial” extensivas àunidades de informação. **Inf. &Soc.:Est.,** João Pessoa, v. 17, n. 1, p. 97-107, jan./abr. 2007. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/viewFile/503/1469>. Acesso em: 16 set. 2014.