USO DA FREQUÊNCIA DE PALAVRAS NA ANÁLISE DA GRAMÁTICA DA INFÂNCIA DE JOÃO RIBEIRO

Márcia Antônia Guedes MOLINA[1]
Docente do Curso Interdisciplinar em Ciência e Tecnologia – UFMA

Arthur Vinicius Sousa SILVA[2]
Aluno do Curso Interdisciplinar em Ciência e Tecnologia – UFMA

Moisés Rocha dos SANTOS[3]
Aluno do Programa de Pós-Graduação em Ciência da Computação – UFMA

RESUMO
O objetivo deste trabalho é avaliar *Grammatica Portugueza (Curso Primário)*, de João Ribeiro, utilizando a análise de frequências de palavras do texto, observando qual a classe de palavras mais valorizada pelo autor, na instância da produção de sua obra. Utilizou-se um programa escrito na linguagem de programação *Python* como suporte para os estudos linguísticos. Analisaram-se, na obra, as partes da classificação, morfologia e sintaxe. O método seguido foi o da análise de conteúdo (BARDIN, 1977). Constatou-se que a classe gramatical em destaque é o verbo, visto que este possui maior frequência de ocorrência e encontra-se de forma bem distribuída por todo o texto.
Palavras-chave: Grammatica Portugueza (Curso primário). João Ribeiro. Análise computacional.

Introdução

Em meados do século XIX, o Brasil passava por um processo de busca de identidade nacional. A língua, como elemento primordial dessa identidade, ensejou vários estudos, debates, publicações. Em especial, os estudos da língua portuguesa no país começavam a ganhar destaque dentro do âmbito intelectual; desde então, muitos pensadores, sobretudo no último quartel do século, começaram a se dedicar a demonstrar as diferenças entre o

[1] Endereço eletrônico: maguemol@yahoo.com.br
[2] Endereço eletrônico: arthsilva6@gmail.com
[3] Endereço eletrônico: moisesrs@ecp.lsdi.ufma.br

português falado no Brasil e em Portugal, favorecendo o início do processo de gramatização brasileira, que produziu tecnologias linguísticas, como dicionários e gramáticas.

Dentre os estudiosos que se destacaram, podemos citar José de Alencar, que, segundo Fávero e Molina (2006), buscou, em determinadas obras românticas brasileiras, justificativas para validar novas regras para falar e escrever o português do Brasil, procurando marcar uma suposta superioridade de nossa língua, dizendo ser ela fruto da evolução da falada em Portugal. Dá-se grande destaque também a Júlio Ribeiro, visto que a publicação de sua obra *Grammatica de Lingua Portugueza* (1881) foi tomada como "divisora de águas", iniciando o período conhecido como científico.

Muitas correntes filosóficas influenciavam os intelectuais brasileiros, como o positivismo e o evolucionismo. Paralelamente, surgiram os primeiros estudos de psicologia, mostrando que a criança, diferentemente do compreendido até então, não era um adulto em miniatura, mas sim um ser com características próprias, necessitando de um material didático voltado especialmente a ela. Dentre as gramáticas destinadas à infância, a *Grammatica Portugueza (Curso Primário)*, de João Ribeiro, ganhou bastante destaque, sendo adotada alguns anos após sua publicação pelos professores do Colégio Pedro II. Temos de nos lembrar que João Ribeiro, embora nascido e formado na égide da gramática de cunho filosófico, pertence ao momento cientificista dos estudos da linguagem, ou seja, momento em que a Gramática deixava de ser vista como arte, passando a ser entendida como ciência.

A obra foi editada pela primeira vez em 1888, corrigida e melhorada diversas vezes, mas sempre voltada ao público infantil. Divide-se em quatro partes, tratando dos sons, da classificação das palavras, das formas (morfologia) e da composição delas no discurso (sintaxe). Em relação à classificação, trata de oito categorias: substantivos, qualificativos, determinativos (onde se incluem os pronomes e os artigos), o verbo, advérbio, preposição, conjunção e interjeição. Na parte da morfologia, observa-se uma vasta importância dada ao verbo, ocupando 30 páginas das 130 da obra infantil. Por fim, na parte da sintaxe (conforme o autor, a parte mais importante da gramática), classifica as proposições em simples e compostas.

A análise de frequência de palavras consiste em contar a ocorrência de uma determinada palavra em um texto para aferir, aliado a outros conhecimentos prévios, a importância deste termo em determinado contexto. A localização de uma palavra no texto pode, atrelado a outros fatores, atribuir a sua importância para o todo. Uma palavra que ocorre pontualmente pode ter importância local ou a uma seção dedicada a esta, enquanto que uma

palavra que ocorre de forma bem distribuída pode ter importância significativa para o texto como um todo.

O objetivo do trabalho é o de avaliar a obra de João Ribeiro, utilizando a análise computacional como suporte para os estudos linguísticos aqui ensejados, observando qual a classe de palavra mais valorizada pelo autor. Para extração destas informações do texto, foi utilizada a linguagem de programação *Python*, devido a esta ser uma linguagem de propósito geral, gratuita e com suporte de fácil manipulação de dados textuais.

Para a análise de frequência de palavras do texto, seguiram-se os seguintes passos: fez-se o pré-processamento do texto para que este pudesse ser analisado computacionalmente; depois, com o auxílio de um programa escrito na linguagem de programação *Python*, foram extraídas a frequência e a localização dos nomes das classes gramaticais; por fim, com os resultados computacionais obtidos, fez-se a interpretação com base nos conhecimentos relacionados às gramáticas da primeira infância. Fica configurado, portanto, um trabalho interdisciplinar, entendendo a interdisciplinaridade como "interação existente entre duas ou mais disciplinas" (FAZENDA, 2008, p. 18). O método utilizado neste trabalho foi o da análise de conteúdo (BARDIN, 1977).

Aparato teórico

História das ideias linguísticas no Brasil

De acordo com Auroux (1989), uma ideia linguística é todo saber construído em torno de uma língua, em um dado momento, como produto quer de uma reflexão metalinguística, quer de uma atividade metalinguística não explícita. De acordo com Fávero e Molina (2006), a história das ideias linguísticas permite estudar as primeiras gramáticas escritas por brasileiros e também qualquer outro saber fundado na ciência linguística, como o estudo, no Brasil, das obras gramaticais surgidas a partir do compêndio de Júlio Ribeiro (1881).

Para a realização deste estudo, é necessário que seja feito o levantamento do maior número possível de fontes para análise; porém, algumas limitações e obstáculos devem ser considerados, dentre eles, temos: a exaustividade, a busca das fontes e o estudo da documentação. Delessalle e Chevalier (1986) afirmam que quanto mais o inventário aumenta, mais esfumaçada a noção de exaustividade, ou melhor, mais seu caráter ilusório e ideológico se afirma. A busca das fontes também apresenta uma série de dificuldades, como o acesso à

documentação, visto que nem sempre é de fácil obtenção, e a seleção do material, em que se considera o fato de que nem sempre é possível localizar as obras. Por fim, o documento deve ser interpretado no contexto em que foi criado. Para Fávero (1996), no estudo da documentação, deve-se considerar a distância espaço-temporal entre o momento em que as obras que constituem o objeto de estudo foram produzidas e o contexto em que se produz o trabalho.

Em suma, em um estudo das ideias linguísticas, além de localizar a fonte de um pensamento, deve-se analisar, no contexto em que determinada ideia foi criada, como foi compreendida, difundida, interpretada e representada, favorecendo uma melhor compreensão da linguística atual. (FÁVERO E MOLINA, 2006)

Análise de conteúdo

Segundo Bardin (1977), a análise de conteúdo é um conjunto de técnicas de análise das comunicações que utiliza procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, sendo uma hermenêutica controlada, baseada na inferência. É um método bastante empírico que depende do tipo de "fala" a que se dedica e do tipo de interpretação que se pretende como objeto, contendo algumas regras de base dificilmente transponíveis. Trata-se de um conjunto de instrumentos metodológicos em constante aperfeiçoamento, que se aplicam a documentos diversificados.

Ainda segundo o autor, a análise de conteúdo possui duas funções: uma heurística e uma de "administração de prova". A primeira "é a análise de conteúdo para ver o que dá", em que o caráter exploratório da pesquisa é enriquecido, incentivando a formulação de hipóteses sobre o objeto estudado, ou como diz o autor, "enriquece a tentativa exploratória e aumenta a propensão para descoberta"; a segunda "é a análise de conteúdo para servir de prova", que busca a confirmação ou a afirmação da hipótese elaborada inicialmente. Na prática, ambas as funções podem coexistir de maneira complementar.

A análise de conteúdo pode ser abordada de maneira quantitativa ou qualitativa, cada qual com seu campo de ação. A quantitativa obtém dados por meios estatísticos, sendo uma análise mais objetiva e exata, uma vez que a observação é mais bem controlada, ou seja, o que serve de informação é a frequência com que surgem certas características no conteúdo analisado; essa análise é útil nas fases de verificação das hipóteses. Já na qualitativa, a presença ou ausência dessas características é o foco do estudo; essa corresponde a um procedimento mais maleável e adaptável à evolução das hipóteses, devendo ser utilizada nas

fases de "lançamento" das mesmas, já que permite seguir possíveis relações entre o índice das mensagens e diversas variáveis das situações de comunicação.

O método de análise de conteúdo proposto por Bardin é dividido em três fases: a pré-análise, a exploração do material e o tratamento dos resultados, a inferência e a interpretação. A pré-análise é a fase que tem por objetivo a organização propriamente dita, sistematizando e tornando operacionais as ideias iniciais; trata-se de estabelecer um programa que deve ser preciso, recorrendo ou não a um computador, de maneira a conduzir a um esquema do desenvolvimento das operações. Esta primeira fase possui, geralmente, três subdivisões: a escola dos documentos que serão submetidos à análise, a formulação das hipóteses e dos objetivos e a elaboração de critérios para a interpretação final.

A segunda fase, exploração do material, é a aplicação sistemática das decisões tomadas, sejam procedimentos aplicados manualmente ou de operações efetuadas por computador. É dividida em três etapas: a escolha das unidades de registro, que visa à categorização e a contagem de frequências; a seleção das regras de enumeração e a categorização.

Por fim, na terceira fase, os resultados são tratados de maneira a serem significantes e válidos, sendo estes submetidos a operações estatísticas simples ou mais complexas, permitindo o estabelecimento de quadros de resultados, diagramas, figuras e gráficos, os quais evidenciam as informações obtidas pela análise.

Data Mining

Conforme Cabena (1997), *Data Mining* (ou mineração de dados) é a técnica de extrair informação, previamente desconhecida e de máxima abrangência, a partir de bases de dados, visando a utilizá-la na tomada de decisão. Uma decisão é baseada, inicialmente, na fonte dos dados, onde há milhões de registros sem um valor agregado. Em outras palavras, são todas as técnicas que permitem extrair conhecimento de uma "massa" de dados. Através dessas técnicas, podem-se desenvolver aplicações que venham a obter informações críticas, com o objetivo de auxiliar o processo decisório de uma organização. Esse processo pode ser aplicado em várias áreas, sendo possível obter vários tipos de descoberta de conhecimento, como associações, agrupamentos, classificações, padrões sequenciais, hierarquias de classificação *etc*.

De acordo com Quoniam *et al* (2001), o processo de *Data Mining* apresenta algumas etapas. Inicialmente, é feita a definição clara do problema que será o objeto de estudo; em

seguida, faz-se a seleção de dados (a partir de uma base de dados bruta), a fim de identificar todas as fontes internas e externas de informação e selecionar um subconjunto de dados necessário para a aplicação do procedimento.

A etapa seguinte é a de preparação dos dados (que inclui o pré-processamento), responsável por 60% do processo e a que exige mais esforço, sendo dividida em ferramentas de visualização e ferramentas de reformatação dos dados. A preparação dos dados é fundamental para a qualidade final dos resultados, o que torna as ferramentas utilizadas muito importantes. Os *softwares* aplicados nesta etapa devem ser capazes de executar diversos processos, como: filtrar variáveis, efetuar conversões, trabalhar com base de dados relacionais, transformar dados em informação útil *etc*.

A última etapa do processo corresponde à análise dos resultados obtidos, possuindo dois aspectos fundamentais: informar novas descobertas e apresentá-las de maneira que possam ser exploradas. Nesta fase, é necessário um especialista na área de estudo em que o processo está sendo aplicado, para que seja feita a devida interpretação dos dados e soluções de questões específicas que possam ser levantadas durante a análise. Em suma, esta é a fase na qual o *Data Mining* é aplicado diretamente e em que ocorre a assimilação de conhecimento.

Procedimentos e métodos

Pré-processamento

Inicialmente, fez-se o escaneamento da obra, delimitando a análise aos campos da classificação, da morfologia e da sintaxe. Utilizou-se a *scanner HP Photosmart C4480*, que possui a funcionalidade de reconhecimento de palavras, o qual facilita no processo de extração do texto. Em seguida, transferiu-se o texto para um formato adequado para o tratamento, sendo escolhido o TXT (Arquivo de Texto); logo após, retiraram-se os caracteres especiais (*, #, -, _, etc) e as pontuações. A linguagem de programação utilizada é *case-sensitive*, que diferencia letras maiúsculas de minúsculas, portanto, optou-se por usar todo o texto em letras minúsculas (ex.: substantivo ≠ Substantivo).

Uso do Python

O *Python* é uma linguagem de programação simples e de excelente funcionalidade para processamento de dados, compatível com todas as plataformas. Possui uma extensa biblioteca padrão, incluindo elementos para programação gráfica, processamento numérico e conectividade na *web*. É próxima a uma linguagem natural e de fácil aprendizado, apresentando boa funcionalidade de manipulação com sequência de caracteres, além de sintaxe e semântica claras. Por ser uma linguagem dinâmica e orientada a objetos, o *Python* permite encapsular e reutilizar dados e métodos; adicionar atributos e digitar variáveis dinamicamente, facilitando o rápido desenvolvimento.

Frequência de palavras

Analisou-se a frequência das palavras correspondentes às classes gramaticais contidas na obra em estudo, sendo elas: substantivo, qualificativo, determinativo, verbo, advérbio, preposição, conjunção e interjeição. Para obter a quantidade de ocorrências mais próximas do total possível, contaram-se tanto as palavras no plural quanto no singular (ex.: quantidade de verbo + quantidade de verbos = total de verbos).

Localização das palavras

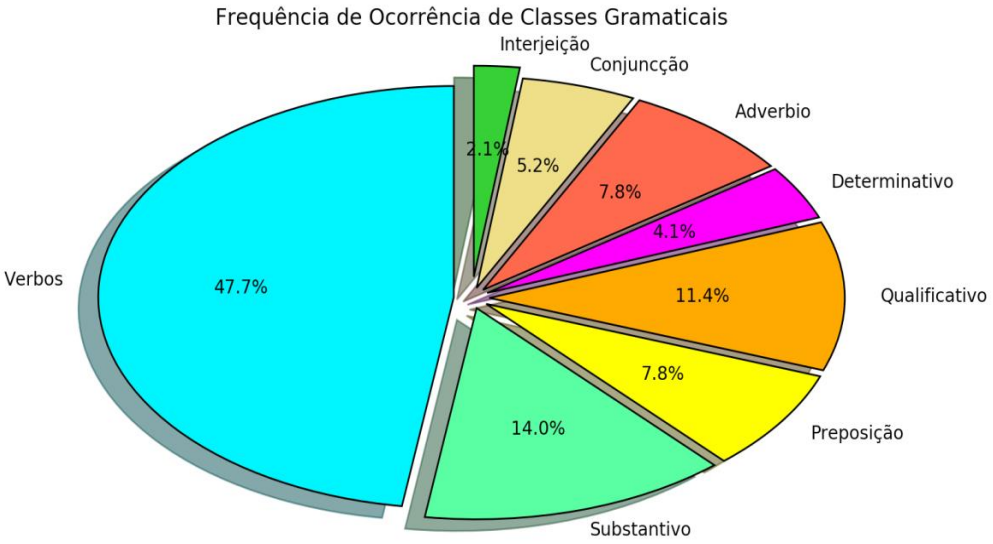
Para a localização das palavras, utilizou-se um programa baseado em *dispersion plot*, da biblioteca NLTK (conjunto de ferramentas de linguagem natural, do inglês NATURAL LANGUAGE TOOLKIT) do *Python*, com o objetivo de obter a disposição dos nomes das classes gramaticais no texto. Optou-se por utilizar o nome das classes somente no singular, visando a detectar quando o autor estivesse falando diretamente de cada classificação.

Resultados e discussões

Após finalizada a análise computacional, os dados referentes a cada classe gramatical foram explorados sob duas perspectivas, quanto a frequência de ocorrência e quanto a distribuição no texto. Para evidenciar as frequências de ocorrência de referências a classes gramaticais, elaborou-se um gráfico-pizza, onde cada setor representa uma dada classificação

e sua porcentagem em relação ao total de citações feitas pelo autor no decorrer da obra. Estas informações estão dispostas no Gráfico 1.

Gráfico 1: Frequência de ocorrência de referências a classes gramaticais

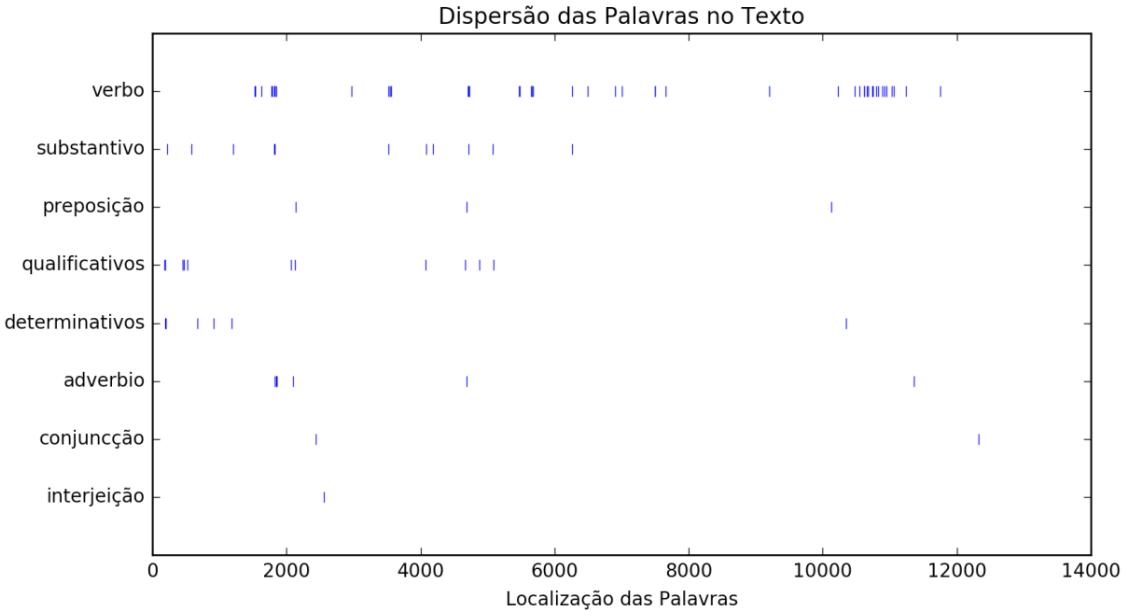


Fonte: Elaborado pelos autores

Com os dados do gráfico, constata-se que a classe mais citada pelo autor é o verbo (47.7%), seguido pelo substantivo (14%), qualificativo (11.4%), advérbio e preposição (7.8%), conjunção (5.2%), determinativo (4.1%) e interjeição (2.1%). Tal resultado demonstra que o autor considera o verbo a palavra por excelência, visto que é a classe gramatical mais abordada durante a obra. Segundo Câmara Jr. (1970), João Ribeiro foi o autor que sistematizou com mais rigor o estudo dos verbos, o que reforça a relevância dos resultados obtidos.

Em seguida, observou-se a distribuição das classes por todo o texto, sendo estas dispostas em um gráfico de dispersão, relacionando o nome de cada classificação com sua localização. No eixo y (vertical) estão as palavras que serão analisadas e no eixo x (horizontal) representa a posição medida em quantidade de ocorrência de palavras. A localização das palavras no decorrer do texto está disposta no Gráfico 2.

Gráfico 2: Distribuição das palavras no texto



Fonte: Elaborado pelos autores

Pode-se perceber que algumas classes são citadas durante toda a obra, enquanto outras ocorrem de maneira isolada. Nota-se que classes como interjeição e conjunção apresentam uma ocorrência pontual, podendo-se concluir que estas são tratadas somente em seções específicas da obra. De fato, a interjeição é citada somente na parte referente às classificações, enquanto que a conjunção volta a ser mencionada na sintaxe, nas seções de concordância e análise da proposição complexa. Diferentemente, o verbo e o substantivo encontram-se bem distribuídos no texto. O substantivo é bem frequente nas partes de classificação e morfologia, porém, praticamente não aparece no estudo da sintaxe.

Novamente, o verbo é a palavra em destaque, sendo bastante citado de forma bem distribuída em toda a gramática. Como foi dito inicialmente, na morfologia, por exemplo, a flexão do verbo ocupa 30 das 130 páginas da obra; o mesmo ocorre para as partes de classificação e sintaxe, em que tal classe é também tratada com ênfase.

Conclusões

Antes de finalizar, recorda-se que o objetivo do trabalho é avaliar, utilizando a análise computacional, qual a classe de palavras mais citada por João Ribeiro. Ao final das análises, pôde-se afirmar que o autor pontuou o verbo com mais destaque em detrimento das demais

classes gramaticais. Recorda-se que o verbo era considerado pelos clássicos como palavra por excelência; para os latinos, enquanto não se pronunciava essa classe, nada podia ser dito.

Embora João Ribeiro fosse um homem que produziu sua obra à luz da gramática científica, como recebeu formação humanística, reflete suas ideias na obra produzida. A importância do verbo para João Ribeiro é pontuada anos depois por Câmara Jr. (1970), quando este afirma que foi ele um dos autores a enfrentar o estudo dessa classe gramatical, apresentando uma partição do verbo em radical e demais elementos.

Referências

AUROUX, Sylvain. *Histoire des idées linguistiques*. Liège/Bélgica: Editions Mardaga, 1989.

BARDIN, Laurence. *Análise de Conteúdo*. Lisboa/Portugal: Edições 70, LDA. 1977.

CABENA, Peter *et al*. *Discovering data mining*: from concept to implementatio. New Jersey: Prentice Hall, 1997.

CÂMARA JR., J. M. *Estrutura da língua portuguesa*. Petrópolis/RJ: Vozes, 1970.

DELESALLE, Simone; CHEVALIER, Jean-Claude. *La linguistique, la grammaire et l'école: 1750-1914*. Paris/França: A. Colin, 1986.

FÁVERO, Leonor Lopes. *As concepções linguísticas no século XVIII*: a gramática portuguesa. Campinas/SP: Editora da UNICAMP, 1996.

FÁVERO, Leonor Lopes; MOLINA, Márcia A. G. *As concepções linguísticas no século XIX*: a gramática no Brasil. Rio de Janeiro: Editora Lucerna, 2006.

FAZENDA, Ivani. *O que é interdisciplinaridade?* São Paulo: Cortez Editora, 2008.

QUONIAM, Luc *et al*. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. *Ciência da Informação*, v. 30, n. 2, p. 20-28, 2001.

RIBEIRO, João. *Gramática portuguesa (Curso Primário).* 94. ed. Rio de Janeiro: Livraria Francisco Alves, 1937 [1888].

RIBEIRO, Júlio. *Grammatica Portugueza*. São Paulo: Jorge Seckler, 1881.

USE OF THE FREQUENCY OF THE WORDS IN THE ANALYSIS OF THE GRAMACY OF THE INFANCY OF JOÃO RIBEIRO

ABSTRACT
The aim of this study is to assess the book "Grammatica Portugueza (Curso Primário)", which was written by João Ribeiro, using a frequency analysis of words from the text. It was observed which word class was the most used by the author in the production of his work. A program written in the Python programming language was used as support for linguistic studies. It was analyzed in the literature parts of the classification, morphology and syntax. For this study the content analysis was the method followed (BARDIN, 1977). It was verified that the grammar class in emphasis is the verb, since this one has a higher frequency of occurrence and it is well distributed throughout the text.
Key words: *Grammatica Portugueza (Curso primário). João Ribeiro. Computational analysis.*