Analisador Léxico-Morfológico de Redações de Estudantes no Estilo do ENEM

Lucas Busatta Galhardi
Universidade Estadual de Londrina
Londrina, Brasil
lucasbgalhardi@uel.br

Cinthyan R. Sachs C. de Barbosa
Universidade Estadual de Londrina
Londrina, Brasil
cinthyan@uel.br

Joao Coelho Neto
Universidade Estadual do Norte do Paraná
Cornélio Procópio, Brasil
joaocoelho@uenp.edu.br

Jacques Duílio Brancher
Universidade Estadual de Londrina
Londrina, Brasil
jacques@uel.br

ABSTRACT

This paper aims to study the students' essays written according to the rules required the ENEM exam (High School National Exam) from Brazil, for the construction of a natural language processor, which initially is concerned with lexical-morphological analysis. The essays were taken from a site which stores texts for students who look for experts' evaluation. To perform the analysis, a system written in Python was developed to achieve the desired results at the lexical-morphological level. The collected information was stored and an interface was created to access it. Variants found for each word is shown as well as some morphological information.

RESUMO

Este trabalho tem como objeto de estudo as redações de estudantes, escritas no mesmo estilo que é exigido no ENEM (Exame Nacional do Ensino Médio), para a construção de um processador de linguagem natural no que tange inicialmente a análise léxico-morfológica. As redações foram retiradas de um site que armazena textos de estudantes que buscam avaliação de especialistas. Para realizar a análise, um sistema escrito em Python foi desenvolvido para alcançar resultados desejados no nível léxico morfológico. As informações coletadas foram armazenadas e uma interface foi criada para acessá-las. Variantes encontradas sobre cada palavra são mostradas, bem como algumas informações morfológicas.

Palavras-chave:
Análise Léxico-morfológica; Redações de Estudantes; Processamento de Linguagem Natural.

Palavras-chave ACM
Natural Language Processing

INTRODUÇÃO

Um léxico ou um dicionário é uma estrutura de dados em que as palavras são armazenadas junto com algumas de suas informações. O léxico é fundamental para formar uma sentença coerente, pois é necessária a compreensão de cada palavra para compreender a sentença [11].

Assim, o léxico é uma lista de palavras que jaz na mente dos falantes. Quando afirmamos isso, estamos deduzindo que, no cérebro dos seres humanos, há um dicionário, um inventário de palavras (p. 131) [16].

Há duas maneiras de aprender uma língua. Uma, natural, por tentativas cada vez mais aperfeiçoadas de comunicação que chegam a conhecimentos memorizados dessa língua (competência natural), como o da criança na família, e, nesse caso, pode-se dominar perfeitamente uma língua sem ser capaz de descrevê-la. A outra, artificial e metalinguística, pela consulta de dois tipos de obras descritas conhecidas como indispensáveis e complementares: a gramática e o dicionário (p. 45) [11].

Quanto ao aspecto morfológico, o importante é identificar a estrutura de uma palavra, como ela é formada e quais são as possíveis variantes de uma palavra específica, mantendo seu significado.

Dentro de um mesmo tipo de palavra, existem grupos de regras que caracterizam o comportamento de um subconjunto de vocábulos da linguagem (exemplo: formação do plural de substantivos terminados em "ão", flexões dos verbos regulares terminados em "ar", etc.) [8]. Assim, a morfologia trata as palavras quanto a sua estrutura, forma, flexão e classificação, no que se refere a cada um dos tipos de palavras.

Portanto, o conceito mais comum sobre Léxico diz respeito a uma lista de itens que existem na língua, os quais um falante precisa conhecer e tem que estocar, por serem signos arbitrários, idiossincráticos e muitas vezes

imprevisíveis – não presumíveis de alguma forma (p. 133) [16].

Ao identificar palavras, a importância está no significado léxico. No processo de inflexão, o significado lexical é preservado e são consideradas formas diferentes da mesma palavra. No entanto, no processo de derivação, uma palavra é modificada e assume um novo significado, construindo uma nova palavra. Distinguir entre esses dois é importante para diferenciar adequadamente as palavras ou agrupá-las para o mesmo significado léxico.

A língua portuguesa pode ser considerada morfologicamente rica porque é uma língua fusional. Tem mais variação nas formas de palavras do que o inglês, que é uma língua analítica, embora tenha menos variação do que uma linguagem aglutinante, tomando a análise morfológica um desafio.

O objetivo deste trabalho é realizar uma análise léxico-morfológica no texto extraído das redações dos estudantes, escritas no mesmo estilo que é exigido no ENEM (*Exame Nacional do Ensino Médio*) realizado no Brasil.

Este artigo foi divido em cinco seções: a primeira seção contextualiza a introdução e a definição do objetivo geral dessa pesquisa; na segunda seção aborda a estrutura de dados utilizada neste trabalho; na terceira seção descreve a construção da arquitetura do sistema para realizar a análise morfológica; na quarta seção apresenta alguns exemplos de uso e considerações do sistema. Finalmente, na quinta seção e última seção, aponta as conclusões e trabalhos futuros.

BASE DE DADOS

Os dados utilizados neste trabalho podem ser encontrados em um *site* que mantém um banco de ensaios chamado Banco de Redações UOL [5]. A cada mês, uma proposta de redação é adicionada ao banco de dados e os alunos podem acessá-lo e escrever suas redações sobre o tema proposto. Cada proposta recebe cerca de dezenas de redações e vinte são escolhidas para serem avaliadas. Essa avaliação é feita por especialistas do proprietário do *site* UOL e segue os mesmos critérios utilizados para avaliar as redações do ENEM.

Até março de 2017, quando iniciamos este audacioso projeto, o banco de dados era composto por 2100 redações e 111 propostas de tópicos. Para facilitar o acesso aos dados e compilação para criar o corpus de redações UOL, [10] desenvolveu um *web crawler* para solicitar as páginas web do UOL e armazenar as propostas e redações em um único arquivo XML. Assim, este *corpus* pode ser usado para avaliar técnicas de Processamento de Linguagem Natural (PLN) e algoritmos de classificação que buscam avaliar corretamente uma redação.

A utilização de Linguagem Natural em conjunto com outras aplicações auxilia quando necessitamos alcançar uma determinada solução/resposta para um problema, trazendo assim benefícios onde a interação do software com o usuário será feita de forma mais simples e um simples usuário que não possua um amplo domínio da linguagem utilizada em um banco de dados, por exemplo, possa interagir facilmente com o sistema (p. 9) [9].

Algumas pesquisas como [2,3,14] realizaram experimentos neste *corpus* de redações para resolver algumas tarefas: avaliação automática de redações seguindo os critérios ENEM, detecção de palavras com erros ortográficos, detecção de tópicos fora de escopo da redação e *feedback* sobre a escrita e nos serviu de base para desenvolver um analisador morfológico da Língua Portuguesa, onde inicialmente vinte redações foram selecionadas do corpus.

O título da proposta de redação é "Direitos em conflito: liberdade de expressão e intimidade". Uma parte de uma redação estudantil sobre essa proposta está no exemplo: "*... No entanto, sempre que possível, a justiça deve prevalecer à liberdade de expressão sobre o direito à privacidade, pois o Brasil é um país democrático, onde todos os cidadãos tem o direito de expor suas opiniões e também serem informados sem manipulações e censura, para que assim a população tenha informações suficientes para tirarem suas próprias conclusões ... ".*

Os dados de trabalho utilizados têm as seguintes estatísticas: 20 redações, 5464 registros no total, 4803 registros removendo pontuação, 1326 registros únicos, 1360 pares únicos (*token*, POS-tag) e 1043 lemas únicos.

ARQUITETURA DO SISTEMA

Como entrada, o sistema recebe um texto bruto para ser processado vindo de um arquivo XML. Como saída, gera duas listas. A primeira é uma lista de todos os diferentes lemas usados no texto. Então, se uma palavra aparecesse de várias formas, somente a forma base estaria nessa lista. Quando uma das palavras da Lista Base é selecionada, todas as suas formas são exibidas.

A segunda saída é uma lista de palavras, com todas as formas aparecidas no texto. Cada palavra tem um registro, no qual algumas informações sobre essa são armazenadas. Os campos de cada palavra são: significado, POS-*tag*, frequência, ocorrências de texto, seu lema, raiz, sufixo, morfemas e o final que indica as inflexões da palavra como número, sexo, pessoa e tempo.

Para alcançar os resultados desejados, o texto foi manipulado principalmente em 9 etapas, em código Python com bibliotecas, que são:

1. **Pré-processamento:** identificação e normalização dos *tokens*, deixando todos em minúsculas e removendo a pontuação. Isso foi feito utilizando a biblioteca python NLTK;

2. Uma lista das **palavras únicas** entre todas as usadas nas redações;
3. A distribuição de **frequência** de cada palavra da lista única, usando o NLTK;
   O **POS-*Tag*** de todas as palavras. Alguns deles tinham mais de uma *tag*, dependendo do contexto. Suas múltiplas *tags* foram mantidas. O *tagger* utilizado nesta fase foi um *tagger* NLTK treinado para Língua Portuguesa [7];
4. A partir da lista de palavras diferentes, nesta fase, uma lista de **lemas** foi criada. O *lema* da palavra é sua forma básica, de onde derivam todas as suas outras variantes. Então, nesse passo, cada palavra foi mapeada para o seu correspondente lema. Isso permitiu construir a primeira saída desejada: as *palavras base* e suas *variantes* encontradas no texto, juntamente com a contagem de ocorrências de cada uma. O processo de lematização foi feito utilizando o LemPORT [13], com uma adaptação para chamar a ferramenta escrita em Java do código Python;
5. Atribuição do **significado** de dicionário das palavras usando uma API JSON pública [1];
6. As palavras foram divididas em seus **morfemas**. Um *morfema* é a menor parte de uma palavra que mantém seu significado. A ferramenta usada para isso foi [15];
7. Morfemas podem ser de três tipos: base, derivada e flexível. *Bases* também chamados de **lexemas ou radicais** são aqueles que contêm o significado lexical. O lexema é o elemento comum em uma família de palavras como comparar em comparação, comparação, comparativo e incomparável. Um Stemmer, da biblioteca NLTK, foi usado para recuperar os lexemas da maioria das palavras;
8. A lista de **ocorrências** para cada palavra. Isso foi feito salvando as posições dos tokens na lista de todos os tokens do texto.

Depois de todos esses passos automáticos também foi feito um trabalho manual que consistiu em: preencher o significado do dicionário quando necessário (a API obteve 2/3 dos casos), confirmando as POS tags e analisando e preenchendo o número, sexo, pessoa, tempo e aspecto, para substantivos, verbos, pronomes e adjetivos, se necessário [4].

ANALISADOR LÉXICO-MORFOLÓGICO

Para fazer uso do sistema, uma interface de linha de comando foi construída. No entanto, para aumentar a experiência do usuário, uma biblioteca foi usada para colorir as entradas e saídas da interface. Essa consiste de menus com opções numeradas para escolher e campos para digitar uma letra ou uma palavra para pesquisa.

Inicialmente, há um menu principal, no qual o usuário pode entrar no menu de lista de bases, lista de palavras ou sair do sistema como visto na Figura 1.

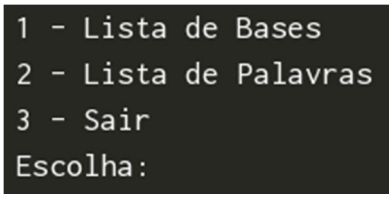

Figura 1. Menu principal.

Fonte: O Autor.

Ao selecionar a primeira opção, outro menu é mostrado (Figura 2) com as opções para ver todas as palavras, ver apenas palavras que começam com uma letra específica, consultar por uma palavra ou voltar. A segunda opção dará um menu semelhante, mas para as palavras em vez das bases. As opções para listar (todas ou por primeira letra) simplesmente exibirão todas as palavras registradas.

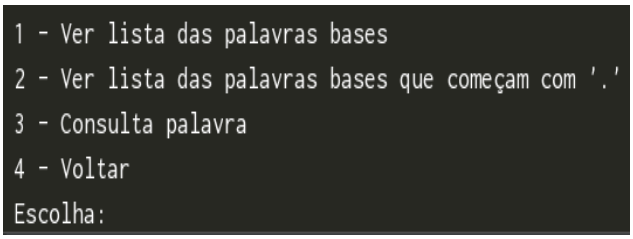

Figura 2. Menu da lista de bases.

Fonte: O Autor.

Na lista de base, quando o usuário seleciona a opção de consultar e digita a palavra, o sistema exibirá todas as formas diferentes que a palavra assumiu no corpus analisado. Como exemplo, na Figura 3 foi consultado o verbo haver, o que resultou em 6 variantes que estão nas redações dos alunos, junto com a frequência de cada uma delas.

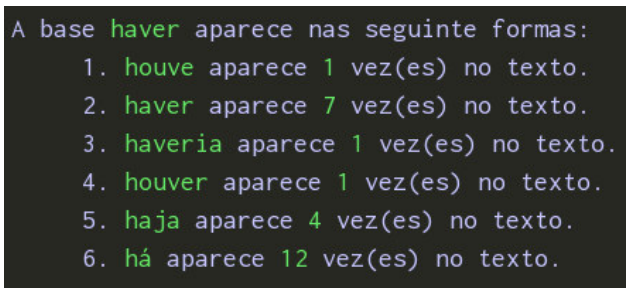

Figura 3. Consulta à palavra "haver".

Fonte: O Autor.

Outro exemplo, de uma palavra muito utilizada, pode ser visto na Figura 4. Nela, é possível ver o resultado da consulta do verbo ser, o que resultou em 18 variantes que estão nas redações dos alunos.

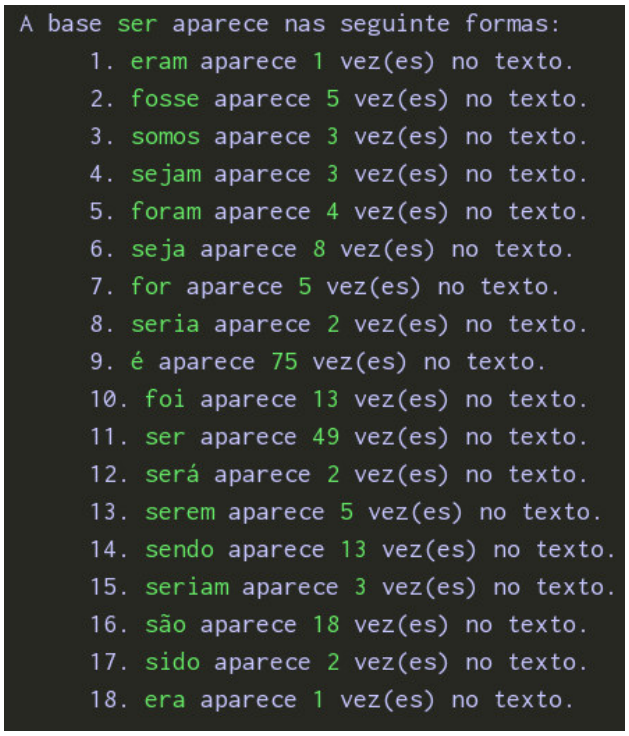

Figura 4. Consulta à palavra "ser".

Fonte: O Autor.

A seguir, na lista de palavras, procurando por uma palavra específica, todas as suas informações são exibidas ao usuário, conforme ilustrado nas figuras 5 e 6.

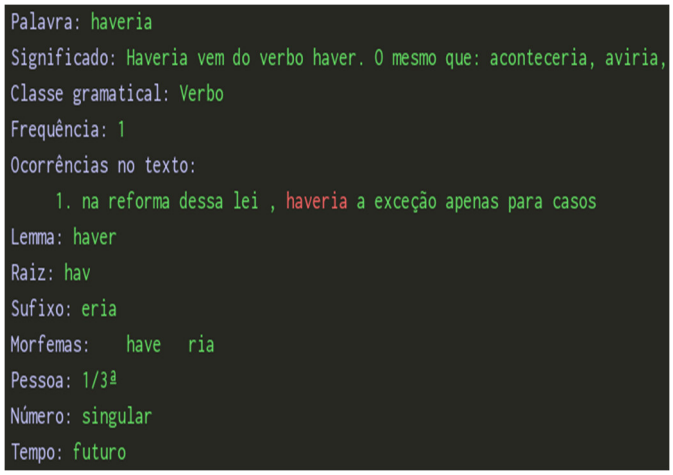

Figura 5. Palavra pesquisada "haveria" e suas informações exibidas.

Fonte: O Autor.

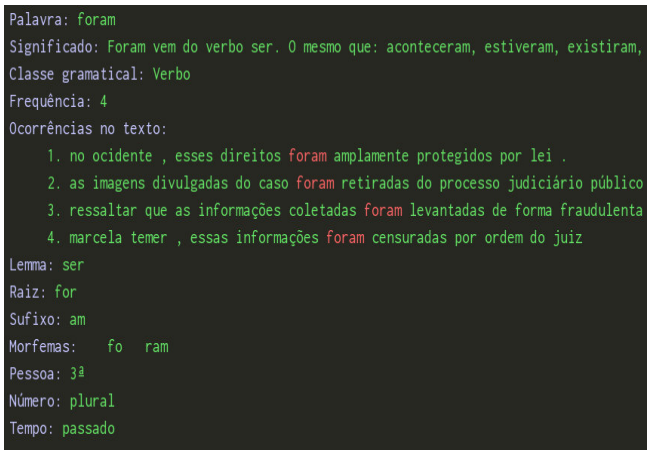

Figura 6. Palavra pesquisada "foram" e suas informações exibidas.

Fonte: O Autor.

5. CONCLUSÕES

Este trabalho abordou uma análise léxico-morfológica da Língua Portuguesa, a qual é necessária como primeira etapa em Sistemas de Processamento de Linguagem Natural. O corpus escolhido para a realização da referida análise consiste de redações escritas por alunos brasileiros, no estilo exigido pelo ENEM.

O título da proposta que as redações trabalharam foi "Direitos em conflito: liberdade de expressão e intimidade" e muitas das palavras que mais apareceram tem relação direta com o tema.

Para a realização da análise foi construído um sistema em Python, que integra algumas bibliotecas bem utilizadas para extrair informações interessantes sobre cada palavra como: significado, classe gramatical, frequência, ocorrência em contexto, forma sem variação morfológica, raiz, sufixo, morfemas, entre outras. Isso demonstra a capacidade que o sistema atingiu ao analisar cada palavra individualmente, fornecendo suas principais características.

O trabalho passou por várias etapas, desde a aquisição dos dados pela internet, sua manipulação para separar as palavras e armazenar cada informação em um registro e, por fim, a criação de uma interface para acesso aos dados. Todo esse processo nos forneceu uma grande visão sobre a riqueza morfológica da Língua Portuguesa, assim como a diversidade de palavras e quantas informações uma porção de texto (20 redações a princípio) possui; como as palavras são utilizadas em suas diferentes formas e as classes gramaticais que cada uma das palavras pode assumir. A interface também forneceu uma maneira simples, confortável e completa para a visualização dos resultados em análise.

Extensões deste trabalho poderiam ser feitas no sentido da observação das palavras dos textos que estavam presentes nos ensaios de redação do ENEM, para ver se essas têm

polaridades positivas ou negativas, como no trabalho de [6], para analisar a correlação de informações léxicas em textos em Português, analisando também características psicológicas dos estudantes.

Como trabalho futuro também é possível continuar a análise desse corpus em outros níveis de análise da Linguagem Natural, como nas análises sintática e semântica dos textos. Outra possibilidade é a criação de uma interface em um sistema *web* para atrair mais usuários a sua utilização, estimulando o conhecimento acerca das palavras e suas variações. Além disso, é possível também realizar a análise da mesma forma em outro corpus e então estudar os resultados de forma comparativa.

REFERÊNCIAS

1. Dicionário Aberto. 2018. *Dicionario Aberto*. Acesso em 28 de Janeiro de 2018 em https://dicionario-aberto.net/search-json/
2. Evelin Amorim e Adriano Veloso. 2017. A Multi-Aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain.
3. Bruno S. Bazelato e Evelin C.F. Amorim. 2013. A Bayesian Classifier to Automatic Correction of Portuguese Essays. Nuevas Ideas en Informática Educativa. In *Anais da XVIII Conferência Internacional sobre Informática na Educação* (TISE'13), Porto Alegre, Brasil, 779-782.
4. Dicio. 2018. *Dicio*. Acesso em 28 de Janeiro de 2018 em https://www.dicio.com.br/
5. UOL Educação. 2018. *Banco de Redações*. Acesso em 28 de Janeiro de 2018 em https://educacao.uol.com.br/bancoderedacoes/
6. Antonio A. A. Machado, Magalí T. Longhi, Maria A. S. N. Nunes e Thiago A. S. Pardo. 2015. Personalitatem Lexicon: Um Léxico em Português Brasileiro para Mineração de traços de Personalidade em Textos. In *Anais do XXVI Simpósio Brasileiro de Informática na Educação* (SBIE'15), 1122-1126.
7. F. Maruki. 2016. *Nltk-Tagger-Portuguese*. Acesso em 28 de Janeiro d 2018 em https://github.com/fmaruki/Nltk-Tagger-Portuguese
8. João Mendes de O. Neto, Sávio D. Tonin, and Soraia S. Prietch. 2010. Processamento de Linguagem Natural e suas Aplicações Computacionais. In *Anais do II Escola Regional de Informática*. SBC.
9. João Mendes de Oliveira Neto, Sávio Duarte Tonin, and Soraia Silva Prietch. 2010. Processamento de Linguagem Natural e suas Aplicações Computacionais. Acesso em 12 setembro de 2018 em https://www.inpa.gov.br/erin2010/Artigo/Artigo9.pdf
10. Guilherme Passero. 2018. *UOL Redações XML*. Acesso em 28 de Janeiro de 2018 em https://github.com/gpassero/uol-redacoes-xml
11. Jesette Rey-Debove. 1984. *Léxico e Dicionário*. Trad. Clívis Barleta de Morais. Alfa. São Paulo.
12. Elaine Rich, Kevin Knight, Pedro Antonio G. Calero, Trescastro Bodega e outros. 1993. *Inteligência Artificial*. São Paulo: Makron Books.
13. Ricardo Rodrigues, Hugo G. Oliveira e Paulo Gomes. 2014. *LemPORT: a high-accuracy cross-platform lemmatizer for portuguese*. OASIcs-OpenAccess Series in Informatics.
14. Jário J. Santos, Ranilson Paiva, Ig I. Bittencourt. 2016. Lexical-Syntactic Evaluation of written activities based on Genetic Algorithm and Natural Language Processing: An experiment on ENEM. Brazilian *Journal of Computers in Education*.
15. Abo Samoor. 2018. *Polyglot*. Acesso em 28 de Janeiro de 2018 em https://github.com/aboSamoor/polyglot
16. Braulino Pereira de Santana. 2013. Morfologia e Léxico atacam as palavras. *Estudos Linguísticos e Literários*, n. 48, jul-dez.