Linguisticun: Uma Ferramenta de Auxílio ao Ensino da Língua Portuguesa e à Linguística Computacional

Rosaine F. Semler[1], Paulo Jr. Varela[1], Michel Albonico[1], Jhonnatan R. Semler[1]

[1]Departamento de Informática – Universidade Tecnológica Federal do Paraná (UTFPR) – Francisco Beltrão – PR – Brasil

rosainefiorio@gmail.com,{paulovarela, michelalbonico, jhonnatanricardo}@utfpr.edu.br

Abstract. *This paper presents a tool for feature extraction, based on structural levels of the Portuguese language. The central idea is to identify the function that each word exerts within a text. The tool offers three levels of feature extraction: (i) word function, such as: subject and predicate; (ii) grammatical levels, which demonstrate the relationship between words in a sentence, such as: nominal and verbal phrases; and, (iii) morphological classes, such as: verb, adjective and adverb. The Linguisticun tools, which is suitable for linguists, experts, researchers and, especially, teachers and students of Portuguese language, is a tool to assist in processes that involve the authorship analysis and supports the teaching-learning process.*

Resumo. *Este artigo apresenta uma ferramenta para extração de características, baseada em níveis estruturais da língua portuguesa. A ideia central é identificar a função que cada palavra exerce dentro de um texto. A ferramenta oferece três níveis de extração de características: (i) a função das palavras, tais como: sujeito e predicado; (ii) níveis gramaticais, que demonstram a relação entre as palavras em uma frase, tais como: sintagmas nominais e verbais; e, (iii) classes morfológicas, tais como: verbo, adjetivo e advérbio. O Linguisticun, propícia à linguistas, peritos, pesquisadores, e, principalmente, professores e alunos de Língua Portuguesa uma ferramenta tanto de auxílio em processos que envolvam a análise de autoria, como de apoio ao processo de ensino-aprendizagem.*

1. Introdução

Com os avanços tecnológicos e a crescente necessidade de aplicações na linguística computacional, torna-se evidente a necessidade de implementação de softwares que possam trabalhar com diferentes recursos linguísticos. Entre as aplicações podem-se citar: a disputa da autoria de textos [Varela *et al* 2018], a identificação de plágios [Sousa-Silva 2013] e a análise de mensagens anônimas [Zheng *et al* 2006][Halvani 2016]. Em língua portuguesa, devido à sua grande variação morfológica e sintática isso se torna um processo complexo. Contudo, a análise linguística é rica em atributos morfológicos, sintáticos e semânticos, e com isso, a identificação destes atributos podem auxiliar na solução desses casos [Smarsaro 2004].

Além disso, com a linguística pode-se vislumbrar novas perspectivas para o ensino da língua portuguesa, tendo como objetivo, a incorporação na prática dos professores da concepção interativa da linguagem. Um outro ponto, é a valorização do sujeito e das

DOI: 10.5753/cbie.sbie.2019.11

variedades linguísticas, entretanto, o que se observa é que a realidade do ensino da língua se mantém baseado na literatura e na gramática na forma escrita da língua [Soares 2001].

Sendo assim, o uso da linguística computacional, que envolve em suas estruturas os conceitos da computação, da estatística e da linguística para manipular a linguagem humana se tornam essenciais [Silva 2006]. No entanto, ainda são ínfimas as soluções computacionais que auxiliam linguistas, peritos, pesquisadores, juristas e estudantes em geral no que concerne a extração de atributos para análises das características em língua portuguesa. Isto abre uma lacuna para o desenvolvimento de uma ferramenta para extrair e manipular informações sintáticas de textos. No entanto, é necessário entender como funcionam as estruturas linguísticas, quer seja, na língua ou sob a visão computacional.

A linguística se refere ao estudo da linguagem humana em seus aspectos práticos, onde utiliza dos conhecimentos adquiridos para melhorar a comunicação pelo uso da língua; e teóricos, onde estuda as características presentes em uma determinada língua [Crystal 2000]. Em consonância, a aplicação de recursos computacionais na linguística é advinda da década de 1950, onde o impulso principal se deu pelo desenvolvimento da inteligência artificial e de programas de tradução automática [Othero 2006]. Observa-se, então, que com os avanços da computação, tornou possível encontrar abordagens diferenciadas para problemas descritivos e práticos que anteriormente não poderiam ser resolvidos. Uma dessa abordagens constitui-se na linguística de corpus, que utiliza de computadores para armazenar e acessar textos escritos ou falados, que sendo legíveis por máquina, podem ser analisados e gerar informações a respeito das particularidades e construções de determinada língua [Othero e Menuzzi 2005]. Um outro foco da linguística computacional é o desenvolvimento de programas capazes de interpretar e gerar informação a partir da linguagem, e são chamados de processadores de linguagem natural - PLN. Dentro desse contexto, alguns softwares propõem-se a reconhecer a categoria que as palavras pertencem - marcadores de categorias gramaticais. Sistemas que realizam a análise estrutural e constituição das frases são denominados analisadores sintáticos [Othero 2006] [Othero e Menuzzi 2005].

Para realizar a análise sintática computacional, o processamento de linguagem natural faz o tratamento de sons, palavras, sentenças e discursos, ou seja, dos aspectos relacionados à comunicação oral ou escrita [Allen 1995]. Em relação a gramática computacional, que é constituída por regras e sentenças que compõem uma língua [Bouillon 1998] tem dupla função: (*i*) a função normativa que define regras para a combinação de palavras com o objetivo de gerar sentenças corretas; e (*ii*) a função representativa que faz a associação entre as frases e a representação sintática. Para tanto, a gramática deve ter uma forma que a represente computacionalmente, e uma dessas formas é a de "*phrase-structure gramar*" que é definida a partir de uma quadrupla <*T, N, P, S*> onde o *T* representa as palavras, o *N* as categorias funcionais e lexicais, o *P* as regras de produção e *S* é um símbolo inicial que pertence a *N* [Gonzalez e Lima 2003].

A análise sintática efetua a análise de como as regras gramaticais são combinadas, de forma a gerar uma árvore que represente as estruturas que constituem a sentença avaliada. Neste artigo, optou-se pela análise sintática, pois em relação a literatura, identificou-se que na linguística computacional existe uma lacuna que permite a construção de ferramentas para extrair e quantificar as características de um texto.

Apesar das evoluções tecnológicas crescentes e das pressões para que a escola, e, consequentemente, ensino agreguem e explorem as tecnologias informação e

comunicação, especificamente, no caso da disciplina de Língua Portuguesa houve pouca evolução no que concerne às práticas de ensino e a inclusão de tecnologias como método enriquecedor da didática. Com isso, através da análise sintática computacional é possível a construção de um software que possibilite ao aluno observar de maneira dinâmica a análise gramatical. Neste caso, mostrando para o aluno as funções das palavras dentro da frase, e, também, que ele possa realizar comparações entre as formas de escrita dos autores de diferentes estilos. Sendo assim, é possível participar da construção do conhecimento do aluno, se tornado agente atuante no processo educacional, deixando de ser agente passivo, de forma que possa passar a refletir e ser crítico de sua língua materna.

A contribuição principal deste artigo é apresentar uma ferramenta computacional para o processo de análise e extração de características em língua portuguesa, baseada nos níveis estruturais da língua, que são: função das palavras, níveis gramaticais e classes morfológicas. Com isso, procura-se evidenciar a importância do uso de recursos que possibilitem representar o estilo de um texto, de um autor ou de um período literário, por exemplo. Por conseguinte, evidencia-se a contribuição e auxílio que a ferramenta propicia a estudantes, professores, linguistas, peritos e pesquisadores na área de linguística computacional em língua portuguesa, mas principalmente, a possibilidade de sua aplicação como ferramenta auxiliar no aprendizado da Língua Portuguesa.

Este artigo está estruturado da seguinte maneira: A seção 1 consiste em apresentar a introdução. A parte 2 descreve os materiais e métodos. Os resultados são apresentados na parte 3. Enquanto, na parte 4 é apresentado um comparativo com a literatura. E, na parte 5 são apresentadas as conclusões e trabalhos futuros.

2. Materiais e Métodos

Nesta seção, são detalhados os materiais e os métodos. Inicialmente foi aplicado um conjunto de questionamentos para peritos, linguistas e professores da Língua Portuguesa para efetuar o levantamento de requisitos, para identificar as principais necessidades e funcionalidades, para definição do escopo da ferramenta. Diante disso, foram desenvolvidos os diagramas para um melhor entendimento da estrutura e funcionalidades. A modelagem escolhida foi a *Unified Modelling Language* - UML. Na área da pesquisa cientifica, este artigo utiliza procedimento de estudo de caso, com seu objetivo exploratório e sua finalidade aplicada.

Na Figura 1 é apresentado o diagrama de caso de uso, que representa as principais funcionalidades do software, que foi baseada nos seguintes requisitos funcionais: (*i*) permitir a seleção de um diretório contendo arquivos do tipo texto (*.txt*); (*ii*) possuir funcionalidade de cadastramento de autores; (*iii*) possibilitar a seleção de características sintáticas que serão extraídas do texto processado; (*iv*) possibilitar a consulta geral das características extraídas por meio de interface gráfica; (*v*) gerar um arquivo com o conjunto de dados escolhido no formato .ARFF (*Attribute-Relation File Format*) para uso em aprendizagem de máquina.

Entretanto, na documentação do software encontrado para download em http://www.utfpr.edu.br/… estão os detalhamentos dos diagramas de classes, máquina de estados, sequência e demais.

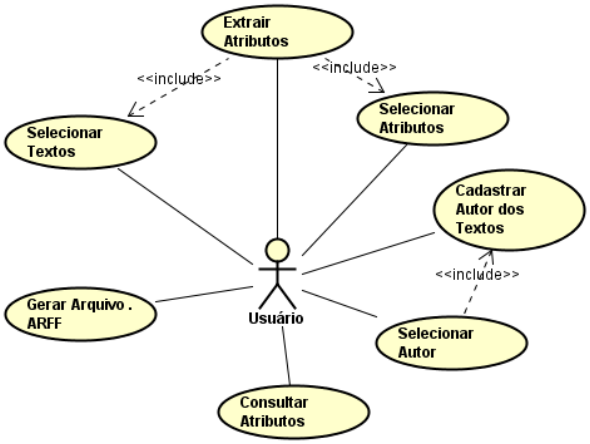

Figura 1 - Diagrama de Caso de Uso

A ideia central é que a ferramenta seja capaz de realizar os processos de acordo com a visão geral (Figura 2). Neste contexto, um ou diversos textos selecionados pelo usuário podem ser inseridos na ferramenta. Diante disso, o usuário pode escolher os níveis de extração: morfológico, sintático ou gramatical, escolhendo quais a características quer usar no processo de extração. Após isso, ocorre o processo de extração, onde a ferramenta atua em conjunto com a biblioteca CoGrOO [Silva 2013]. Por fim, é efetuada a normalização do modelo e gerado o conjunto de dados para uso em ferramentas de aprendizagem de máquina. Demais detalhes são apresentados na seção 3.

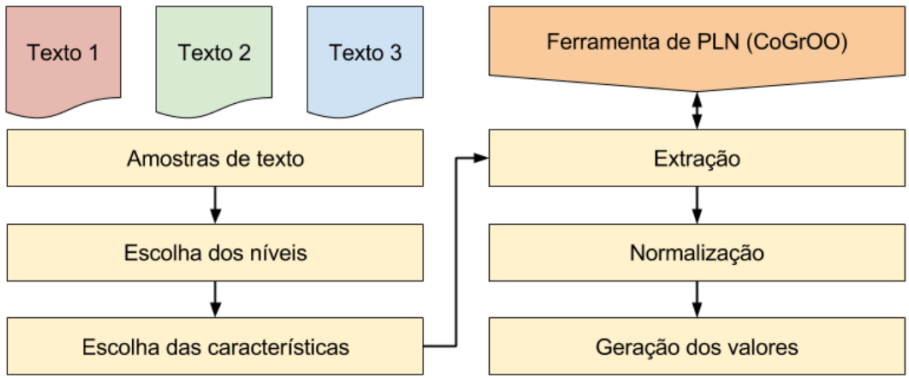

Figura 2 - Visão Geral do Processo

Com a estrutura da ferramenta planejada, ilustrada e documentada por meio dos diagramas, foi então realizado o levantamento das tecnologias que atendem às propostas de desenvolvimento. Optou-se pela utilização de tecnologias de livre utilização, como a linguagem de programação Java e banco de dados MySQL.

Posteriormente, foi iniciado o processo de implementação do software, que considera a análise sintática de textos a sua parte principal. Sendo assim, foi realizado levantamento sobre *frameworks* que dispusessem de funções de análise sintática em língua portuguesa. Diante disso, foi escolhida a biblioteca CoGrOO [Silva 2013], a qual apresenta as funções de rotulagem necessárias para identificação das funções sintáticas, viabilizando assim o desenvolvimento da ferramenta.

Após o desenvolvimento, realizou-se uma fase de testes. Nesta fase, um *corpus* de textos foi submetido ao processo de extração de características, com a finalidade de avaliar se as extrações estavam ocorrendo de forma correta. Com isso, foi possível efetuar algumas alterações e correções na ferramenta, antes de disponibilizá-la.

3. Resultados

Como resultados, apresenta-se a ferramenta *Linguisticun*. Ao acessar o sistema é apresentado a tela principal, que contém as funcionalidades organizadas em menus. São apresentadas as opções de cadastros, processos de extração e geração de arquivos de conjunto de dados. Na tela de cadastro do autor, é possível realizar o cadastramento, alterações ou mesmo exclusão. Na tela de execução, é exibido a seleção para escolha de arquivos de texto, onde também é indicado o autor ao qual os textos que devem ser processados pertencem, bem como, as características linguísticas a serem extraídas.

O primeiro passo é indicar qual amostra de texto ou conjunto de amostras será submetido ao processo de extração de características. Para a seleção dos textos dos autores de diferentes períodos literários, está disponível um corpus de textos para que seja possível que o professor navegue pela complementaridade dos conteúdos de estilos literários e gramática no ensino de Língua Portuguesa. Quando submetido ao processo de extração, o texto é atribuído à um determinado autor conhecido ou não. Posteriormente, são realizadas as configurações para a extração das características, que no caso o professor poderá orientar em qual nível se baseará o estudo. De acordo com a seleção das características, são gerados os vetores com os valores correspondentes de forma normalizada. O processo de extração é customizado e foi desenvolvido uma interface gráfica que permite a extração baseada nas características da língua portuguesa, conforme a Figura 3. Todas as informações resultantes das análises de textos, além dos registros de processamento de textos e autores, são armazenadas em banco de dados. E, na sequência, o aluno verifica as características extraídas acessando a tela de consulta.

A ferramenta foi desenvolvida de forma a possibilitar a extração de 3 níveis de características sintáticas em separado ou em conjunto. Quando se opta pelo 1º nível é possível fazer a extração de características baseadas em informações sintáticas, tais como: sujeito, verbo, predicado e complemento. Tais características também podem ser extraídas em separado ou em conjunto. Um exemplo é dado pela análise da frase: “Acreditamos sempre em dias melhores”, onde o sujeito oculto é revelado a partir do número apresentado pelo verbo acreditar (acreditamos); o predicado verbal é designado por meio da ação que fica implícita no sujeito (acreditamos sempre em dias melhores); é identificado um adjunto adverbial de tempo expresso pela palavra sempre; o objeto direto, que é o complemento do verbo acreditar (em dias melhores); e um predicativo do objetos dada pela palavra “melhores”, que confere uma característica ao objeto direto. Na Figura 4, pode ser vista uma árvore sintática da análise da frase.

Já no 2º nível de extração de características são possíveis de extrair características da gramática sintagmática. Nestas classes estão incluídos os sintagmas nominais e verbais, que apresentam meios mais simplificados para a descrição da estrutura das orações. Os sintagmas são identificados a partir dos elementos que compõem a oração e, devido a sua organização em torno do núcleo da oração mantém a dependência e ordem entre si, sendo que o núcleo sozinho pode constituir o sintagma [Duarte 2012]. Para melhor identificar um sintagma em uma frase, utilizamos o exemplo: “A virtude é uma

característica humana". Quando o núcleo é um nome ou pronome ("a virtude"), tem-se um sintagma nominal. Os demais elementos da oração possuem como elemento fundamental o verbo ("é uma característica humana"), razão pela qual são denominados de sintagmas verbais [Duarte 2012]. Na Figura 5, é possível observar a estrutura dos elementos dos sintagmas através da representação da árvore sintática sintagmática. Percebe-se que existem alguns elementos constituintes da frase, que são essenciais para a análise sintática sintagmática, tais como, o determinante, que é representado geralmente pelos artigos, numerais e pronomes. Existe também o elemento modificador que muitas vezes são representados por advérbios e adjetivos.

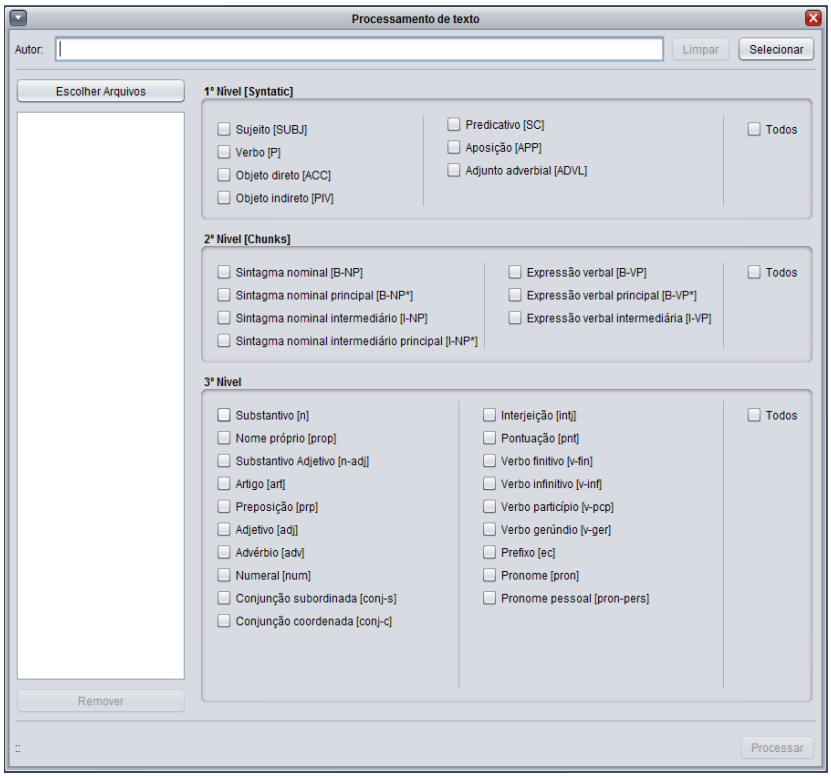

Figura 3 - Tela de Seleção de Características

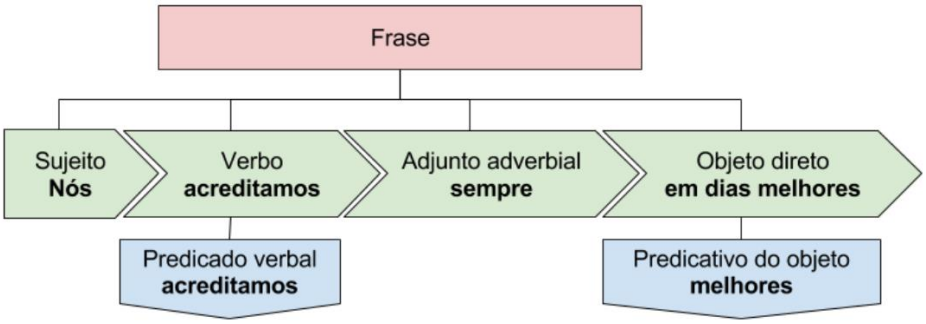

Figura 4 - Exemplo de Geração de Árvore para Extração

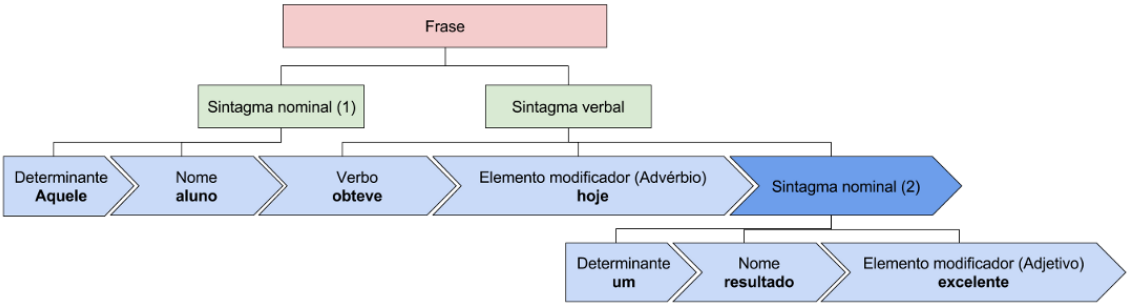

Figura 5 - Exemplo de Árvore Sintagmática

No 3º nível podem ser selecionadas as classes mais comuns das palavras, tais como: substantivos, adjetivos e advérbios. Neste caso, podem-se extrair cada uma das classes de palavras em separado, combinadas ou no conjunto completo. Esta análise é denominada morfológica, pois estuda as palavras de acordo com a classe gramatical a que elas pertencem, como exemplo da Figura 6.

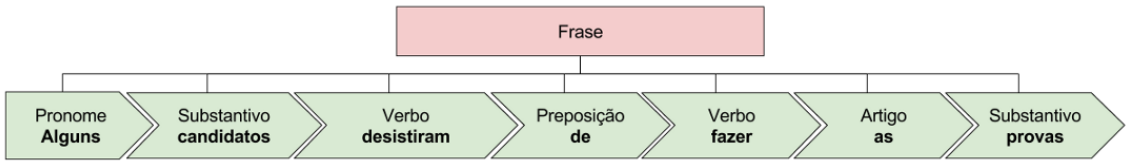

Figura 6 - Exemplo de Árvore Morfológica

Com esses resultados e o processamento de diversos autores, o software permite que o aluno observe dentro de textos de diferentes períodos literários e consequentemente seus autores, as particularidades de escrita de cada autor e período literário que o caracteriza. Além disso, observar de forma gráfica as funções das palavras dentro de uma frase, ou seja, de forma dinâmica, a análise gramatical. Falando mais especificamente, dos estilos literários com a aplicação dos quantitativos, é possível gerar gráficos que demonstrem as diferenças de escritas de cada estilo literário com relação a utilização dos recursos da língua portuguesa.

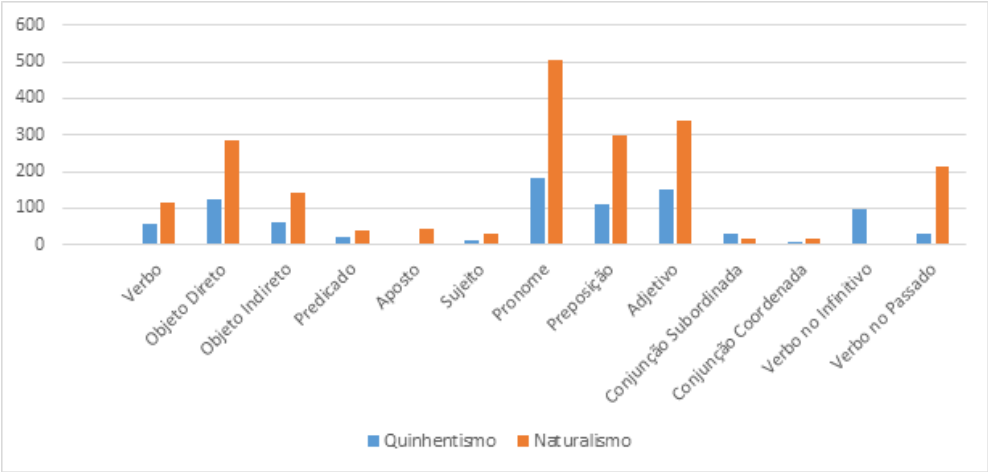

Figura 7 - Comparação entre o uso das características entre dois estilos literários

Observando a Figura 7, infere-se que o quinhentismo utiliza-se de grande quantidade de adjetivos em sua escrita em comparação às demais características, o que se justifica por ser um período de escrita basicamente informativa, ou seja, os navegantes

informando a Coroa de Portugal como era a sua nova Colônia. Já no naturalismo a característica, que mais se sobressai é o pronome, que vai de encontro há duas características do período, que são o objetivismo e a impessoalidade, ou seja, faz uso de pronomes de forma a não deixar explícito o sujeito da frase.

Como opção, a ferramenta gera vetores, que são utilizados em pesquisas de estilometria e determinação de autoria, de acordo com as características selecionadas pelo usuário. Quando os vetores de características de cada texto ou de cada autor são gerados, estes passam por um processo de normalização, que consiste na transformação da frequência absoluta em frequência relativa. A frequência absoluta apresenta o número vezes que uma característica aparece no texto. Já a frequência relativa, indica a representatividade de uma determinada característica em função da quantidade total de palavras do texto. São gerados vetores personalizados para cada experimento e aplicação, podendo este ser ter uma ou várias posições, que são escolhidas e fixadas no software de extração de características.

4. Comparativo dos Resultados com a Literatura

Com o intuito de comparar a ferramenta apresentada neste artigo, realizou-se um estudo comparativo com a literatura. Entretanto, existem diferentes ferramentas e bibliotecas com foco no processamento de linguagem natural, com objetivos diversos e suporte à várias linguagens, conforme demonstra a Tabela 1.

Tabela 1. Comparação de Frameworks e Ferramentas de PLN

Nome da Ferramenta	**Ferramenta / Framework**	**Suporte a Língua portuguesa Brasileira**	**Análise Semântica**	**Análise Sintática**	**Análise Morfológica**
OpenNLP	Framework		X	X	X
UIMA	Framework	X	X	X	X
NLTK	Framework	X	X	X	X
Stanford CoreNLP	Ferramenta		X	X	X
LX-Center	Ferramenta		X	X	X
LingPipe	Framework	X	X	X	X
VISL	Ferramenta	X	X	X	X
ExatoLP	Ferramenta	X	X	X	X
Linguisticun	**Ferramenta**	**X**		**X**	**X**

Observa-se que dentre os *frameworks* e ferramentas abordadas, que todas são *open-source*, ou seja, são de livre utilização. No entanto, existem três *frameworks* que

possuem suporte à Língua Portuguesa Brasileira, no entanto, precisam ser integrados à algum software para que possam ser utilizados em sua integralidade, o que pode dificultar a sua utilização por pessoas que não são da área. Quanto às ferramentas completas, somente o VISL [Bick 2003] e o ExatoLP [Finatto *et al.* 2015] tem suporte à Língua Portuguesa Brasileira. Contudo, permitem a marcação das funcionalidades das palavras em formato de árvore léxica, com marcação por cores diretamente na tela do navegador, não permitindo a geração de arquivos específicos para utilização em softwares, tais como o Weka, utilizado em aprendizagem de máquina. O ExatoLP, necessita de instalação prévia e trabalha com domínios de interesse específico, isto é, uma ferramenta automática que recebe um corpus literário específico, extrai os sintagmas nominais e gera uma hierarquia de conceitos do corpus a partir de abordagens linguísticas e estatísticas.

Percebe-se, então, que a ferramenta *Linguisticun*, permite tanto a marcação das palavras de um corpus, de acordo com os níveis gramaticais escolhidos pelo usuário, além de possibilitar a criação do arquivo específico para utilização em pesquisas relacionadas à aprendizagem de máquina. Sendo assim, um agente facilitador no processo de análise linguística utilizando meios computacionais em língua portuguesa.

5. Conclusões

A linguística computacional é uma das áreas de pesquisa e desenvolvimento que existem diversas lacunas, onde a informatização se faz essencial. Então, é de suma importância desenvolver ferramentas que auxilie em qualquer um dos processos do ensino, da aprendizagem e da aplicação da linguística computacional. A ferramenta Linguisticun permite a extração e análise de características sintáticas da língua portuguesa, e possui a capacidade de extrair três níveis de informações de uma frase, que são: morfológicas, sintagmas e funções sintáticas.

Quanto ao processo de extração de características dos textos, perceberam-se taxas de precisão aceitáveis, no momento da rotulagem de cada palavra. Isso se deve a biblioteca CoGrOO, que é integrada, permitindo uma análise de textos e a extração das características sintáticas de forma mais eficiente. Entretanto, ainda é possível encontrar eventuais erros em resultados de análises, porém, continua sendo uma ferramenta útil para agilizar o processo de extração de características.

Então, fica evidente que apesar da falta de incentivos e da quase inexistência de ferramentas computacionais que auxiliem o professor de língua portuguesa no ensino dos conteúdos de gramática e estilos literários, é importante, o uso de algum tipo de apoio que possa demonstrar aos alunos as diferentes características gramaticais da língua. Além do mais, incentivar a leitura, e a análise das diferentes formas de utilização dos recursos da língua pelos autores da literatura brasileira, em busca de fomentar as discussões entre aluno-professor e aluno-aluno.

A ferramenta foi testada, disponibilizada para a comunidade, e continua em aprimoramento. Em uma próxima versão propõe-se trabalhar com uma arquitetura orientada à serviços. Pretende-se trabalhar com outras línguas, tal como a língua inglesa e a espanhola. Também será disponibilizado um corpus de textos de autores de diferentes estilos da língua portuguesa, e inserido a extração de atributos semânticos. Além disso, realizar uma pesquisa com os usuários do sistema (professores e alunos) para verificar os benefícios e limitações da ferramenta no processo de ensino-aprendizagem.

Referências

Allen, J. (1995) “Natural Language Understanding”. Redwood City, CA: The Benjamin/Cummings Pub. Co., 654 p

Bick, E. (2003). "A Constraint Grammar Based Question-Answering System for Portuguese". In: Fernando Moura Pires & Salvador (eds.) Progress in Artificial Intelligence (Proceedings of EPIA'2003, Beja, Dec. 2003), pp. 414-418. Springer

Bouillon, P. (1998) “Le traitement automatique des langues”. Bruxelles, Duculot, 245p.

Crystal, D. (2000) “Dicionário de Linguística e Fonética”. 8ª Ed. Rio de Janeiro: Ed. Jorge Zahar.

Finatto, M. J. B. Lopes, L. Ciulla, A. (2015) “Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida.”. DOMÍNIOS DE LINGU@GEM, v. 9, n. 5

Gonzalez, M.; Lima, V. L. S. (2003) “Recuperação de Informação e Processamento da Linguagem Natural”. XXIII Congresso da Sociedade Brasileira de Computação, Campinas. Anais do III Jornada de Mini-Cursos de Inteligência Artificial, Volume III, p.347-395

Halvani, O. Winter, C. Plug, A. (2016) “Authorship verification for different languages, genres and topics”. Digital Investigation, vol. 16, pp. 33–43.

Othero, G. A.; Menuzzi, S. M. (2005) “Lingüística Computacional: teoria & prática”. São Paulo: Parábola, 126 p

Othero, G. A. (2008) “Lingüista “puro” vs. lingüista “computacional”: revisitando a distinção entre “lingüista de poltrona” e “lingüista aplicado””. Domínios de Lingu@gem – Revista Eletrônica de Lingüística. Ano 2, nº 1.

Silva, B. C. D. (2006) “O estudo Linguistico-Computacional da Linguagem”. Letras de Hoje. Porto Alegre. v. 41, nº 2, p. 103-138.

Soares, M. (2001) “Que professores de português queremos formar?” Movimento, Niterói, n. 3, p. 149-155.

Smarsaro, A. das D. (2004) “Descrição e formalização de palavras compostas do português do Brasil para elaboração de um dicionário eletrônico”. 130 f. Tese (Doutorado em Letras) - Programa de Pós-graduação em Letras do Departamento de Letras da Pontifícia Universidade Católica, Rio de Janeiro, 2004.

Sousa-Silva, R. (2013). “Detecting plagiarism in the forensic linguistics turn (Ph.D.)”. Aston University

Varela, P. J. Justino, E. J. R. Bortolozzi, F. and Albonico, M. (2018) “A Computational Approach for Authorship Attribution on Multiple Languages”. In 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro.

Zheng, Rong, Jiexun Li, Hsinchun Chen and Zan Huang. (2006) “A framework for authorship identification of online messages: Writing-style features and classification techniques.” JASIST 57: 378-393.

Silva, W. D. C. M (2013) “Aprimorando o COrretor Gramatical CoGrOO”. Dissertação Instituto de Matemática e Estatística da Universidade de São Paulo.