Texto Livre: Linguagem e Tecnologia
E-ISSN: 1983-3652
revista@textolivre.org
Universidade Federal de Minas Gerais
Brasil

Fonseca, Cláudia Aparecida; de Souza Netto, Rafael Santiago; Nascimento Bodolay, Adriana; Carvalho Guelpeli, Marcus Vinícius
AnoTex: rotina de filtragem de dados estruturados do gênero artigo científico como contribuição para o PLN
Texto Livre: Linguagem e Tecnologia, vol. 11, núm. 3, september-december, 2018, pp. 40-64
Universidade Federal de Minas Gerais

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

ANOTEX: ROTINA DE FILTRAGEM DE DADOS ESTRUTURADOS DO GÊNERO ARTIGO CIENTÍFICO COMO CONTRIBUIÇÃO PARA O PLN

ANOTEX: STRUCTURED DATA FILTERING ROUTINE OF THE SCIENTIFIC ARTICLE GENRE AS CONTRIBUTION TO PLN

Cláudia Aparecida Fonseca
Universidade Federal dos Vales do Jequitinhonha e mucuri
claudia.fonseca@ufvjm.edu.br

Rafael Santiago de Souza Netto
Pesquisador do Grupo MTPLNAM
rafael.santiago@oi.com.br

Adriana Nascimento Bodolay
Universidade Federal dos Vales do Jequitinhonha e mucuri
adriana.bodolay@ufvjm.edu.br

Marcus Vinícius Carvalho Guelpeli
Universidade Federal dos Vales do Jequitinhonha e mucuri
marcus.guelpeli@ufvjm.edu.br

RESUMO: A diversidade dos recursos de linguagem, que possibilita a construção de aplicações em Processamento de Linguagem Natural, provoca a necessidade da criação de ferramentas que sejam igualmente flexíveis. Além disso, essas ferramentas devem ser tão amigáveis quanto úteis, a fim de reduzir o esforço para usuários iniciantes e, ao mesmo tempo, promover um eficiente desempenho para usuários avançados. O presente artigo apresenta o AnoTex, que é um anotador textual capaz de executar a filtragem de dados estruturados do gênero artigo científico, coletados dos arquivos disponíveis na base de dados da Biblioteca Eletrônica SciELO – *Scientific Electronic Library On-line*. Como produto do processo de extração, obteve-se uma base de dados com as informações filtradas e estruturadas no formato XML, que delimitam e identificam as marcações do gênero em análise, disponível para uso em várias ferramentas e aplicações. São apresentadas outras ferramentas de anotação de textos, atualmente existentes, e argumenta-se que o AnoTex é o primeiro a combinar um bom nível de facilidade de uso com recursos estruturados, constitutivos do gênero, de alta qualidade linguística. Os resultados demonstram como a categorização dos elementos constitutivos do gênero, por meio de sua representação em bancos de árvore, pode condensar as informações disponíveis de forma hierarquizada e dinâmica, construídas durante a compilação. Essas características podem indicar novas estratégias de uso para as marcações coletadas, de modo a atender às necessidades no melhoramento do acesso e da recuperação da informação proporcionados pelo uso das ferramentas de processamento de texto.
PALAVRAS-CHAVE: Processamento de Linguagem Natural; gênero textual; anotador textual; anotação de corpus.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

ABSTRACT: The diversity of language resources, which enables the construction of applications in Natural Language Processing, causes the need to create tools that are equally flexible. In addition, these tools should be as user-friendly as useful, in order to reduce the effort for new users and at the same time promote efficient performance for expert users. This article presents the AnoTex, which is a textual annotator capable of performing the filtering of structured data of the textual genre scientific article, collected from the available archives in the database of SciELO – Scientific Electronic Library Online. As a product of the extraction process, we have obtained a database structured in the XML format that delimit and identify the markings of the genre under analysis, available for use in various tools and applications. Other textual annotation tools are currently available, and it is argued that AnoTex is the first to combine a good level of ease-of-use with structured, basic text-based features of high linguistic quality. The results demonstrate how the categorization of the constituent elements of the genre, through its representation in tree banks, can concentrate the information available in a hierarchical and dynamic way. These features may indicate new usage strategies for the collected tags to meet the needs for improvement in the access and retrieval of information through the use of word processing tools.
KEYWORDS: Natural Language Processing; textual genre; textual annotator; annotation of corpus.

1 Introdução

Este artigo insere-se no contexto do Processamento de Linguagem Natural (PLN), em que é assinalada uma estreita relação entre anotação e geração de *corpus* com a análise dos elementos constitutivos do gênero artigo científico. Sua necessidade surge do crescente desenvolvimento das tecnologias e a excessiva quantidade de informações disponibilizadas pelos meios de comunicação *on-line*, que gera grandes desafios às pessoas que têm dificuldades para a localização e seleção das informações de seu interesse.

Tal necessidade tem provocado uma mudança de paradigma na forma de fazer pesquisa nas áreas da ciência da linguagem e ciência da computação aqui relacionadas. Essas mudanças podem trazer grandes benefícios, não só para as esferas educacionais, mas também para o cidadão comum nas práticas de recuperação, interpretação e/ou geração textual por meio da criação de modelos computacionais destinados ao PLN. Os modelos computacionais podem ser utilizados tanto para alcançar os propósitos científicos, pois exploram a natureza linguística da comunicação, como também para alcançar propósitos práticos, uma vez que permitem a comunicação entre homens e máquinas (CAMBRIA; WHITE, 2014). Além disso, por meio dos modelos computacionais, é possível perceber e testar o desempenho de algorítimos que tentam simular a compreensão e a produção da linguagem. Nesse processo, quando é evidenciado algum êxito ou falha é possível levantar questionamentos e hipóteses sobre o funcionamento da linguagem que podem ser explorados e, portanto, convertidos em melhorias para os algoritmos de processamento de texto.

Por considerar o assunto sobre processamento de texto de muita relevância, neste artigo é apresentado um modelo computacional para a área de PLN, com interface com a

Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Linguística de Corpus (LC), sustentado pela compilação e anotação de metadados em *corpus,* utilizando conceitos da Linguística Textual (LT), com a finalidade de demonstrar que essa é uma associação possível e necessária, que pode trazer bons resultados para o desenvolvimento das pesquisas em PLN.

O objetivo deste trabalho é sistematizar, por meio do levantamento dos traços característicos do artigo científico, uma base de dados com informações extraídas e estruturadas em formato XML, que possam ser disponibilizadas para várias ferramentas e aplicações, com a finalidade de servir de "modelo didático" para uso de ferramentas computacionais. A constituição desse modelo segue a concepção dinâmica de gênero defendida por Marcuschi (2005, p. 21), inspirado em Bakhtin (1992, p. 279). Para os autores, um gênero textual se caracteriza como formas de enunciados, com padrões relativamente estáveis. Esses enunciados, por sua vez, têm conteúdo temático, estilo e construção composicional constituídos historicamente pelo trabalho linguístico dos sujeitos nas diferentes esferas e na diversidade da atividade humana. Constituição essa que cumpre determinadas finalidades em determinadas circunstâncias, típicas da comunicação em um dado meio social. Todo esse dinamismo confere o estatuto privilegiado para o estudo e a organização dos diversos campos da ciência que utilizam os gêneros discursivos como base de suas pesquisas.

A compreensão textual por meio da leitura, de um modo geral, é essencial para o aprendizado e o desenvolvimento intelectual do aluno, pois lhe permite a busca por novos conhecimentos. Especialmente, a leitura de artigos científicos, no ensino superior, tem implicações diretas na formação acadêmica e no desempenho do discente, pois contribui diretamente com a sua aprendizagem. Sendo assim, o desenvolvimento de ferramentas de Tecnologia Digital de Informação (TDI) que possam facilitar a seleção, recuperação e/ou leitura de textos (em quantidade e velocidade), vem sendo pesquisadas com frequência e otimismo (GAMBHIR; GUPTA, 2017).

Além disso, a importância da leitura no âmbito educacional tem sido objeto de estudo realizado tanto por educadores quanto pesquisadores. Muitos desses estudos, segundo Santos (2015, p. 78), destacam a prática da leitura "como um dos caminhos que levam o aluno ao acesso e à produção do conhecimento, enfatizando a leitura crítica como forma de recuperar todas as informações acumuladas historicamente e de utilizá-las de forma eficiente." Isso significa que, no âmbito educacional, é um dever da universidade fomentar pesquisas que possam trazer boas estratégias capazes de estimular os discentes a desenvolver o interesse pela leitura, desenvolver a capacidade intelectual, científica e dar condições estruturais para a prática da produção científica.

Nesse sentido, argumenta-se que o desenvolvimento de capacidades desse caráter no contexto universitário, em que a produção científica é frequentemente solicitada, é de grande relevância na formação acadêmica do aluno. Acredita-se que a busca por melhorias nos trabalhos com atividades de textualização e retextualização, que surgem da necessidade da produção e transformação de um novo texto, a partir de um ou mais textos-base (MATENCIO, 2002), implica a seleção das principais características da macroestrutura do texto-base de acordo com os propósitos do retextualizador. Sendo assim, em função da excessiva massa de documentos disponível para consulta e leitura no ambiente virtual torna-se importante o uso de ferramentas computacionais que possam auxiliar o usuário na seleção de seus objetivos de leitura.

Algumas estratégias de retextualização, em função do dinamismo, pragmatismo e

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

funcionalidade característicos do gênero textual artigo científico, podem ser exploradas pela PLN, que tem como finalidade o oferecimento de ferramentas e recursos automáticos que permitam ao usuário e/ou produtor de textos:

- acesso a informações *on-line*;
- melhor percepção da mensagem principal – ou o que for mais relevante – em relação ao texto-base, de modo a manter-se atualizado;
- seleção de material relevante para uma leitura eficiente;
- simplificação de textos para atender às necessidades do usuário/leitor, possibilitando ampliação do acesso;
- democratização do acesso à leitura e possibilidade de uma verdadeira inclusão de novos leitores;

Dessa forma, implementar e disponibilizar um modelo computacional que auxilie um estudante ou um pesquisador a compilar e explorar uma coleção de artigos científicos, via uma representação de banco de árvore[1], que expresse características relevantes dos textos e suas relações de similaridade, pode ser muito útil. Essa representação deve permitir a visualização e recuperação das principais características do contexto de produção e da arquitetura geral do texto-base por meio de suas etiquetas de marcação em XML. Esse modelo de linguagem de marcação permite descrever qualquer tipo de dado e, por possuir um padrão aberto para interoperabilidade e intercâmbio de informações, pode ser aproveitado por uma ampla variedade de tecnologias, o que favorece a reutilização tecnológica e facilita a extração de dados de *corpora*.

Neste artigo, são apresentados alguns anotadores de *corpus,* atualmente disponíveis, e argumenta-se que esses não descrevem suficientemente as características do gênero textual em combinação com um nível adequado de facilidade de uso. Em seguida, é apresentado o Anotador Textual – AnoTex, que faz parte de um projeto maior do Grupo de Pesquisa em Mineração de Textos, Processamento de Linguagem Natural e Aprendizado de Máquina (MTPLNAM)[2] da Universidade Federal do Vales do Jequitinhonha e Mucuri (UFVJM). E, por fim, são apresentadas as contribuições e potencialidades deste estudo para a área de processamento de texto.

2 Fundamentação teórica

O PLN é uma área interdisciplinar baseada em um campo de estudo versátil, incluindo a Engenharia Computacional, que fornece métodos para ilustração do modelo, algoritmo e realização; a Linguística, que categoriza formas e práticas linguísticas; a Matemática, que fornece modelos e métodos formais; a Psicologia, que estuda modelos e teorias que motivam o comportamento humano; a Estatística, que oferece procedimentos para prever medidas com base em registros de amostra; e a Biologia, que percorre em

1 Adaptado do termo em inglês, *treebank*, que são *corpora* de dados linguísticos transcritos, enriquecidos com anotações de informações sintáticas e/ou semânticas, na forma de representação arbóreas em que se indicam as relações entre elementos no interior das sentenças ou fragmentos de sentenças (FARIA; GALVES, 2016, p. 300).

2 Grupo de Pesquisa em Mineração de Textos, Processamento de Linguagem Natural e Aprendizado de Máquina. Disponível em: <http://mtplnam.com.br/>. Acesso em 12 jan. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

torno da arquitetura subjacente dos processos linguísticos no cérebro humano (MANARIS, 1998).

Os sistemas de geração de linguagem natural convertem informação de bancos de dados de computadores em linguagem compreensível ao ser humano e sistemas de compreensão de linguagem natural transformam ocorrências de linguagem humana em representações mais formais, mais facilmente manipuláveis por programas de computador[3]. Sendo assim, Da Silva (2006) alerta que os precursores dos estudos em PLN já indicavam que um computador não poderia emular a linguagem humana, satisfatoriamente, se não conseguisse compreender o contexto do assunto em discussão. Seria necessário, para isso, fornecer ao programa um modelo detalhado do domínio específico do discurso em questão. Esse argumento converge com os objetivos da presente pesquisa, cujo princípio acredita na observação da anotação do gênero textual como característica relevante na implementação de sistemas de PLN. Para corroborar, Da Silva (2006) diz ser fundamental que:

> para emular aspectos de língua natural pressupõe equipar um sistema de PLN com vários sistemas de conhecimento e fazê-lo emular uma série de atividades cognitivas: - possuir um “modelo simples de sua própria mentalidade”; - possuir um “modelo detalhado do domínio específico do discurso”; - possuir um modelo que represente “informações morfológicas, sintáticas, semânticas, contextuais e do conhecimento de mundo físico”; - “compreender o assunto que está em discussão”; - “lembrar, discutir, executar seus planos e ações”; - participar de um diálogo, respondendo, com ações e frases, às frases digitadas pelo usuário; - solicitar esclarecimentos quando seus programas heurísticos não conseguirem compreender uma frase (DA SILVA, 2006, p. 122).

Desde seu surgimento na década de 1950, as pesquisas em PLN têm se dedicado principalmente às tarefas de tradução automática, recuperação da informação, sumarização de textos, modelagem de tópicos e, mais recentemente, mineração de opiniões (CAMBRIA; WHITE, 2014). De um modo geral, o PLN preocupa-se diretamente com o estudo da linguagem voltado para a construção de *softwares*. Sendo assim, para a sua implementação são necessários vários subsistemas complexos para representar os diversos aspectos da linguagem, como: sons, palavras, sentenças e discurso nos mais variados níveis estruturais, de significado e de uso (VIEIRA; LIMA, 2001). Mesmo com muitos avanços na área, até o momento, não há disponível para uso um *software* que seja capaz de combinar todas as abordagens e de gerar uma base de conhecimento, que armazene informações que descrevam os recursos de linguagem, necessários para o desenvolvimento de aplicações de PLN, eficientemente. Isso se deve à complexidade, à riqueza de detalhes e às variações da linguagem humana, que dificilmente são adequadas à formalização do computador.

O estudo do PLN é um domínio de pesquisa amplo e fecundo, uma vez que a construção do corpo de conhecimentos necessários para a implementação de sistemas de processamento de texto, com o grau de sofisticação delineado neste trabalho, exige seleção, organização, representação e codificação de uma variedade de informações na complexa tarefa de criar um simulacro computacional da competência e do desempenho

3 Adaptado de <https://pt.wikipedia.org/wiki/Processamento_de_linguagem_natural>. Acesso em: 27 nov. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

linguísticos humanos. Para tanto, é importante esclarecer alguns conceitos e expressões que serão abordados neste trabalho. A anotação de *corpus*, como recurso metodológico, já é um procedimento consolidado no meio científico. Entretanto, a possibilidade de se obter resultados eficientes, com a codificação da variedade de elementos necessários ao processamento de texto, ainda enfrenta alguns desafios. Então, é necessário conhecer os métodos e técnicas mais bem-sucedidas que pesquisadores como Luhn (1958), Vieira e Lima (2001), Da Silva (2006), Aluísio e Almeida (2006), Alencar (2010, 2013a), Paixão (2014), Paixão de Souza, Kepler e Faria (2010, 2014), Souza, Faria e Temponi (2016), Rocha e Guelpeli (2017) demonstram em uma série de trabalhos já desenvolvidos para o PLN.

Por meio dos referidos trabalhos, é possível inferir que uma base de conhecimento é constituída por recursos de linguagem que são ferramentas e dados necessários para a gestão e o desenvolvimento de aplicações de PLN. Esses recursos de dados, que descrevem conhecimento sobre entidades e fatos, possuem descrição e anotação necessárias para o estudo de fenômenos linguísticos, avaliação de sistemas e para o treino e teste de componentes de aprendizagem de máquina utilizados no PLN. Os *corpora* são um recurso de dados constituído por um conjunto de *corpus* com suas anotações de classes gramaticais, morfossintáticas, categorias semânticas, entre outras. Esses recursos são coletados criteriosamente, com o propósito de servirem ao pesquisador ou usuário como meio de exploração da linguagem por intermédio de evidências empíricas, que são extraídas com o auxílio de ferramentas computacionais. *Corpora* anotados linguisticamente se tornaram um recurso importante para pesquisas tanto em linguística como em linguística computacional, uma vez que ambas podem explorar as relações da linguagem em aplicações tecnológicas, tornando melhor a interação entre o homem e a máquina. As ferramentas, por sua vez, são os recursos como os segmentadores de sentenças e palavras, etiquetadores morfossintáticos, recursos de análise léxico-semântica e anotadores textuais, que possibilitam a construção de aplicações de PLN mediante o reuso dos dados.

Dentre essas formas de recursos, este trabalho propõe reutilizar os dados do gênero textual artigo científico coletados da base de dados estruturados da SciELO. Segundo Hovy, Navigli, Ponzetto (2013), recursos estruturados são dados legíveis por máquinas e que codificam relações de vários tipos de acordo com o nível de informação que se deseja marcar. Um desses recursos é a Linguagem de Marcação Extensível (XML)[4], que utiliza etiquetas (*tags* em inglês) para delimitar e identificar as marcações. Essas etiquetas correspondem a recursos estruturados de alta qualidade linguística, pois são anotadas com a ajuda do conhecimento de especialistas de domínio, lexicógrafos e linguistas. A XML é uma linguagem que permite descrever qualquer tipo de dado e é um padrão aberto para interoperabilidade e intercâmbio de informações. Entretanto, não é tarefa fácil encontrar recursos estruturados à disposição do pesquisador, pois sua produção demanda grandes esforços na criação, armazenamento e atualização.

4 Sigla de *eXtensible Markup Language* (Linguagem de Marcação Extensível). Conjunto de regras baseado em *SGML* para codificação de documentos textuais de maneira legível para seres humanos e máquinas, desenvolvido pelo *W3C* (*World Wide Web Consortium*) (GUIA DE USO DE ELEMENTOS E ATRIBUTOS XML PARA DOCUMENTOS QUE SEGUEM A IMPLEMENTAÇÃO SCIELO PUBLISHING SCHEMA, 2016).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Com base na premissa de trabalhos já realizados, quanto à abordagem, as técnicas de PLN podem ser classificadas em três classes, a saber: *estatística*, *linguística* e *híbrida* (BHARTI; BABU, 2017). Dentre as principais técnicas e abordagens de PLN que são direcionadas para o processamento de texto, ou seja, que podem ser usadas em anotadores de *corpus,* é exibido um detalhamento das principais no Quadro 1:

Quadro 1: Principais técnicas em PLN.

Técnica	Abordagem	Descrição	Trabalhos
Remoção de ***Stopwords*** – (Filtragem de Palavras de Parada)	Linguística	Consiste em um processo de filtragem para remoção de palavras de pouca relevância como artigos, preposições, pronomes, conjunções, advérbios, numerais e interjeições.	Luhn (1958).
TF-IDF – (Frequência de Termo - Frequência de Documento Inverso)	Estatística	O *Term Frequency* (TF) baseia-se no pressuposto de que o peso de um termo que ocorre em um documento é diretamente proporcional à sua frequência. *Inverse Document Frequency* (IDF) baseia-se no pressuposto de que a especificidade de um termo pode ser medida por uma função inversa do número de documentos em que ocorre.	Luhn (1958); Rocha; Guelpeli (2017).
Latent Semantic Analysis (LSA) (Análise semântica latente)	Híbrida	Consiste em um método, que utiliza a sinonímia e a polissemia, para extração e representação do significado semântico de palavras em um contexto. Essa representação é obtida por meio de cálculos e aplicações matemáticas que analisam a relação entre termos e documentos, decompondo-os em vetor de índice.	Landauer; Laham (1998).
N-grams	Estatística	Essa técnica consiste na coocorrência de palavras e permite fazer uma predição estatística de dois, ou mais, termos de um texto aparecerem em certa sequência.	Cohen (1995); Alencar, (2010, 2013b).
Segmentation - (Segmentação de texto em frases)	Híbrida	Consiste na segmentação do conteúdo do texto em sentenças individualizadas, representativas de um conjunto semântico mínimo para definição de uma proposição.	Paixão de Sousa; Kepler; Faria (2010); Alencar (2013b).
Tokenization - (Segmentação de texto em palavras)	Híbrida	Consiste no processo que segmenta uma sequência de caracteres do texto em uma sequência de unidades de significado (palavras) que compõem o texto. Os espaços e a pontuação são geralmente adotados como *tokens* delimitadores para idiomas ocidentais.	Webster (1992); Paixão de Sousa; Kepler; Faria (2010); Alencar (2013b).
Stemming - (Lematização e radicalização)	Linguística	A Lematização consiste na representação de cada palavra do texto de entrada em sua forma primitiva (*lemma*). O processo de radicalização das palavras tem como finalidade a remoção de sufixos e prefixos de um termo, para que este seja reduzido ao seu radical (*stem*).	Lovins (1968); Paixão de Sousa; Kepler; Faria (2010).
Part-Of-	Linguística	Consiste em etiquetar as palavras do texto de	Paixão de

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Speech (POS) Tagging - (Etiquetagem morfossintática)		entrada com suas respectivas classes gramaticais e distribuições sintáticas. Essa tarefa atribui uma *tag* (etiqueta), a cada palavra da sentença, que corresponde a sua respectiva classificação.	Sousa; Kepler; Faria (2010); Alencar (2013b); Santos; Zadrozny (2014)

Fonte: Elaborado pela autora.

Todas essas técnicas em PLN têm aplicações variadas nas áreas de resposta automática a perguntas, recuperação de informação, sumarização ou tradução automática de textos, classificação de textos, geração de dicionários, análise de sentimentos, dentre outras (FIALHO et al. 2016). Portanto, necessitam de sistemas eficientes para classificação e reconhecimento para o maior número de elementos textuais possíveis, uma vez que ora utilizam métodos estatísticos, ora utilizam métodos linguísticos e ora utilizam ambos.

3 O cenário da pesquisa

A SciELO – *Scientific Electronic Library Online*[5] (Biblioteca Científica Eletrônica *On-line)* é uma biblioteca virtual de revistas científicas em formato eletrônico. Sua principal finalidade é organizar e publicar textos completos de revistas nacionais dos países da América Latina, Caribe, Espanha, Portugal e, mais recentemente, da África do Sul, na *Web*, assim como produzir e publicar indicadores de seus usos e impactos (PACKER, 1998; CANALES, 2017). O modelo SciELO nasceu com o propósito estratégico de contribuir para o avanço da pesquisa científica gerada em países iberolatinoamericanos, assim como melhorar a qualidade de seus periódicos e aumentar a sua visibilidade, acessibilidade, uso e impacto.

Essa biblioteca opera com a Metodologia SciELO, que inclui critérios de avaliação de periódicos baseados em padrões internacionais, desenvolvida para a preparação, armazenamento, disseminação e avaliação de publicações de comunicações científicas em formato aberto e eletrônico (PACKER, 1998; CANALES, 2017). O projeto, iniciado em 1997, conta atualmente com uma grande rede de informações científicas que, segundo Canales (2017), indexa mais de 1.440 revistas de todas as áreas de conhecimentos e dá acesso a mais de 700.000 artigos. Estima-se um aumento de mais de 40.000 artigos, em média por ano, e a rede de acesso aberto recebe, em média, mais de 1,5 milhão de *downloads* por dia (CANALES, 2017, p. 213). Isso evidencia o significativo impacto que a rede SciELO tem na comunidade científica e acadêmica.

Optou-se pelos artigos científicos da biblioteca SciELO como fonte de dados para a criação do *corpus* de análise, por suas publicações possuírem uma estruturação em linguagem XML, facilitarem o armazenamento em banco de dados e possibilitarem um melhor uso de seus elementos textuais identificados pelas etiquetas de anotação por ferramentas computacionais. Todos os elementos constitutivos do gênero artigo científico recebem uma marcação, como, por exemplo, a *tag* para identificar o autor é <author>, a

5 Disponível em: <http://www.scielo.org>. Acesso em 05 de abr. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

que identifica o título é <title>, dentre outras. O Guia de uso de elementos e atributos XML para documentos que seguem a implementação SciELO Publishing Schema[6] dispõe de mais de uma centena de etiquetas para delimitar os mais variados elementos constitutivos do texto, como: referências bibliográficas, seções, resumo, parágrafos, tabelas, figuras, financiamento, entre muitas outras.

Em função do impacto que as coleções SciELO têm na comunidade científica e acadêmica, instaura-se, assim, a necessidade de alguma interdisciplinaridade, seja teórica ou metodológica, entre os Estudos do Texto e do Discurso, da Linguística de *Corpus* e da Informática. No processo de compilação do *corpus*, buscou-se proporcionar, partindo da descrição e da caracterização – ou mesmo da problematização do estatuto de gênero, diferentes tipos de aplicações para os ambientes educacionais. A partir dos textos estruturados em base de dados, pode ser possível a produção de novas estratégias de uso para essas marcações, de modo a atender às necessidades no melhoramento do acesso e da recuperação da informação, proporcionados pelo uso das ferramentas de PLN.

4 Trabalhos correlatos

Alguns dos principais anotadores textuais, disponíveis para análise, fazem uso de abordagens complexas para a categorização de seus recursos de dados e exploração de correspondência para seus padrões. Dentre elas, as mais utilizadas são: a abordagem *morfossintática* que especifica o modo como os grupos de elementos devem se organizar; e, a abordagem *semântica,* que especifica o que o grupo de elementos deve significar. Segundo Cambria e White (2014), trabalhos mais recentes reconhecem a necessidade de categorizar os recursos de linguagem que expressam o conhecimento externo do texto na interpretação e resposta à entrada de linguagem. Esse conhecimento está relacionado à abordagem *pragmática* que especifica como as informações contextuais podem ser aproveitadas para fornecer melhor correlação entre as variadas abordagens combinadas entre si.

A seguir, serão apresentadas as ferramentas computacionais eDictor, Aelius e COMEDI para anotação e processamento de texto, que utilizam abordagens morfossintática e semântica.

4.1 eDictor

O eDictor[7] é uma ferramenta utilizada para auxiliar a edição eletrônica em XML de textos antigos para fins de análise filológica e a codificação linguística automática (PAIXÃO DE SOUZA, 2014). Essa ferramenta foi idealizada para a criação do Corpus Anotado do Português Tycho Brahe[8] (CTB). Esse editor apresenta as principais funcionalidades:

6 Disponível em: <http://docs.scielo.org/projects/scielo-publishing-schema/pt_BR/1.5-branch/>. Acesso em 05 de abr. 2018.

7 Disponível em: <https://humanidadesdigitais.org/edictor/>. Acesso em: 19 de abr. 2018.

8 Disponível em: <http://www.tycho.iel.unicamp.br/corpus/>. Acesso em: 19 de abr. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

- flexibilidade dos formatos gerados, permitindo tanto a leitura humana como a leitura automática;
- garantia da qualidade filológica da edição por se tratar de um editor especializado;
- possibilidade de operar com vários níveis de edição: Junção, Segmentação, Grafia, Modernização, Expansão, Correção, Pontuação;
- possibilidade de criar novos níveis de edição de acordo com a necessidade do pesquisador.

Dentre seus vários níveis de edição, destacam-se suas principais funcionalidades, que são: junção e segmentação. A primeira é utilizada para unir trechos do texto como palavras quebradas, enquanto a segunda faz o oposto, separa trechos indevidamente unidos. Tais edições são feitas com anotação XML de forma a manter o texto original disponível para consulta. Essas edições nos textos podem incluir outros níveis de trabalho como modernização, expansão, grafia e pontuação. Ao processar um texto longo, ou vários textos provenientes da mesma origem, existe uma grande chance do anotador se deparar com a necessidade de repetir muitas vezes a mesma edição, aumentando a possibilidade de cometer um engano. Para minimizar essa inconsistência, foi idealizado o *software* E-DICMATIC, que, acoplado ao eDictor e por intermédio de algoritmos de aprendizado de máquina, pode ler *corpora* previamente anotados e aprender as edições para aplicá-las em novos textos. Quanto à confiabilidade, este editor mostrou que a codificação em XML com intervenção direta sobre o documento é demasiadamente sujeita a falhas e demanda extensa e incessante revisão da codificação por um editor usuário (humano) PAIXÃO DE SOUSA; KEPLER; FARIA, 2010, 2014) e (SOUZA; FARIA; TEMPONI, 2016).

4.2 Aelius

O etiquetador Aelius[9] é um *software* livre em desenvolvimento, cujo projeto de código aberto tem como objetivo desenvolver um conjunto de módulos baseados em Python, biblioteca *Natural Language Toolkit* (NLTK), e interfaces para ferramentas externas livremente disponíveis para análise superficial do Português Brasileiro. Esse anotador faz parte do projeto *Aelius Brazilian Portuguese POSTagger*[10] e está registrado no *SourceForge.net*[11] (ALENCAR, 2010, 2013a, 2013b). Por possuir uma arquitetura híbrida, recorre às abordagens baseadas em regras formuladas manualmente e ao sistema estatístico estocástico baseado em n-gramas. Esse *software* pretendia, para Alencar (2010):

> satisfazer algumas carências da "linguística de *corpus* do português de orientação diacrônica, ao mesmo tempo em que contribui para sanar a escassez de *corpora* e *language models* do português no NLTK e acrescenta a essa biblioteca algumas funções bastante úteis para o desenvolvimento de etiquetadores mais eficazes (ALENCAR, 2010, p. 2).

9 Disponível em: <http://aelius.sourceforge.net/>. Acesso em 19 abr. 2018.

10 Disponível em: Aelius Brazilian Portuguese Pos-tagger <http://sourceforge.net/projects/aelius/files/>. Acesso em 19 abr. 2018.

11 Disponível em: Sourcefoge.Net - Maior hospedagem mundial de software de código aberto. <http://sourceforge.net/>. Acesso em 19 abr. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

O Aelius foi projetado para etiquetar morfologicamente textos escritos de maneira automática. Para tanto, esse editor desempenha as seguintes tarefas:

- pré-processamento de *corpora*;
- construção de *language models* e etiquetadores com base num *corpus* anotado;
- avaliação do desempenho de um etiquetador;
- comparação entre diferentes anotações de um texto;
- realização de anotação de *corpora* e auxílio na revisão humana de anotação automática.

Essa ferramenta também inclui recursos de idioma, como modelos de linguagem, textos de amostra e padrões de ouro[12]. Atualmente, a Aelius já oferece recursos para *corpora* e corporação de *POS,* e gera anotações em diferentes formatos, como em XML no esquema de codificação TEI[13] P5. Todavia, é necessário depurar e rever o módulo para aumentar a qualidade na etiquetagem, quando o *corpus* é constituído por textos atuais, o que exige uma soma de esforços no treinamento do etiquetador com textos corrigidos manualmente (DA COSTA CARVALHO; VASCONCELOS; DE ALENCAR, 2012).

4.3 COMEDI

O *COmponent Metadata EDItor* (COMEDI)[14] é um editor de componentes para metadados baseado na *Web*, em conformidade com qualquer perfil CMDI[15], e que oferece suporte atualizado para recursos adotado pela CLARIN[16]. No COMEDI, é possível criar um registro de metadados a partir do zero, ou carregar, editar e baixar qualquer arquivo XML CMDI. Seus componentes podem ser usados independentemente do editor, ou podem ser usados por meio de uma interface *web*, em que o usuário seleciona um perfil CMDI para iniciar, e o editor exibe o perfil como um formulário *on-line* simples que oculta o código XML. Além disso, o editor pode funcionar como um servidor completo para armazenar, pesquisar, visualizar e gerenciar, por meio da administração de grupos e de usuários, o controle sobre o direito de acesso aos metadados individuais. O gerenciamento de usuários do editor é feito pela autenticação do *login* e opera em dois níveis: o de usuário e o de grupo (LYSE; MEURER; DE SMEDT, 2014).

A instância do desenvolvedor do COMEDI está atualmente integrada em um

12 Em PLN padrão de ouro significa que as avaliações mais fortes e significativas são baseadas em resultados do mundo real, em que um sistema é implantado operacionalmente e é medido seu impacto nos resultados gerados por usuários do mundo real (REITER, 2018).

13 Text Encoding and Interchange (TEI) define um conjunto vasto de elementos e atributos na linguagem XML que permitem representar características estruturais, conceituais e de visualização dos textos.

14 Disponível em: <http://clarino.uib.no/comedi/page?page-id=repository-main-page>. Acesso em: 19 abr. 2018.

15 Component MetaData Infrastructure (CMDI), fornece uma estrutura para descrever e reutilizar um conjunto de metadados. Disponível em: <https://www.clarin.eu/content/component-metadata>. Acesso em: 19 abr. 2018.

16 Common Language Resources and Technology Infrastructure (CLARIN), é uma infra-estrutura de pesquisa que foi iniciada a partir da visão de que todos os recursos e ferramentas de linguagem digital de toda a Europa e além são acessíveis através de um ambiente on-line de logon único para o apoio de pesquisadores nas ciências humanas e sociais. Disponível em: <https://www.clarin.eu/>. Acesso em: 19 abr. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

framework web no emergente centro nacional CLARIN tipo B da Universidade de Bergen, na Noruega. Um perfil CMDI consiste em *componentes* e *elementos*. Os elementos são os nós terminais, aos quais são atribuídos um valor, digitado em um campo de texto do formulário, ou selecionado em um menu suspenso. Os elementos que pertencem aos terminais, geralmente, são agrupados em componentes, por exemplo, um componente pessoa (com elementos como sobrenome, nome) ou um componente licença (com elementos como nome da licença, URL da licença). Em resumo, os perfis recomendados, e que são encontrados no menu suspenso COMEDI, são os seguintes:

- *corpusProfile:* descreve *corpora* de todos os tipos e modalidades;
- *lexicalProfile:* descreve recursos lexicais;
- O usuário é livre para desenvolver o seu próprio perfil, mas nesse caso, deve reutilizar os componentes existentes, na medida do possível.

Essas recomendações têm como objetivo auxiliar os pesquisadores, individuais ou em grupo, a descreverem eficientemente os recursos linguísticos com metadados de acordo com a estrutura do CMDI. Para isso, o editor precisa distinguir campos obrigatórios de campos opcionais e impor que as entradas obrigatórias sejam preenchidas (DIMA, et al., 2012). Entretanto, um problema com a implementação desse editor é como atribuir direitos de edição de forma segura, para evitar a substituição descontrolada de informações existentes que podem ser compartilhadas por vários usuários. Esses problemas estão relacionados à qualidade dos metadados, especialmente quanto à variação de seus valores, o que tem sido foco em trabalhos de monitoramento dessa natureza. Ainda não é possível sustentar um nível suficiente de atividade concertada para abordar sistematicamente esse problema (KING, et al., 2016).

Mesmo com o avanço nas pesquisas e o desenvolvimento de algoritmos cada vez mais poderosos, ainda não se tem uma ferramenta de processamento de texto com desempenho eficiente. Essas ferramentas se revelam limitadas e com desempenho ruim se não forem devidamente treinadas ou se os contextos e domínios mudarem. Em seguida, será apresentado um modelo computacional que, por meio da abordagem pragmática, filtra e categoriza os recursos de dados contextuais do artigo científico, demonstrando a relevância desse tipo de abordagem para a compreensão e categorização de padrões do contexto de produção desse gênero.

5 Modelo computacional – AnoTex

O AnoTex[17] é um anotador textual em desenvolvimento, cuja finalidade é extrair as principais características constitutivas do gênero artigo científico. Sua principal atribuição é processar os elementos do contexto de produção e da infraestrutura geral do texto, para a compilação do *Corpus* de Artigos Científicos (CorpACE). São filtrados da estrutura básica do gênero em XML, definido pelo nó raiz (<article>), os principais elementos

17 Ferramenta computacional versão 0.1 beta, linha de comando, escrita em ANSI C, desenvolvida para funcionar em qualquer sistema que disponha de uma libc. Até o momento, foi usada no Linux, nos principais BSDs (FreeBSD, NetBSD e OpenBSD) e no Windows, desenvolvida pelo programador do Laboratório de Pesquisa MTPLNAM – Rafael Santiago de Souza Netto. <http://lattes.cnpq.br/4425978431209527>.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

distribuídos em suas três filiações dos nós: <front> constituído de elementos pré-textuais, <body> constituído de elementos textuais e <back> constituído de elementos pós-textuais. Dos elementos pré-textuais do <front>, que caracterizam e descrevem os metadados do artigo, são filtrados: título, autoria, afiliação, resumo e palavras-chave. Dos elementos textuais do <body>, que caracterizam e descrevem o corpo textual do artigo, são filtradas as seções. Cada uma delas possui um elemento <title> seguido de um ou mais parágrafos <p>. Seções de primeiro nível podem ser qualificadas de acordo com seu tipo por meio do atributo <sec-type>. Dos elementos pós-textuais do <back>, que caracterizam e descrevem a parte final do artigo, é filtrada a lista de referências. Além disso, são filtrados da estrutura básica do gênero em PDF os elementos de destaque em negrito e itálico. Todos esses dados são coletados e exportados dos arquivos extraídos do banco de dados da SciELO.

O conjunto de requisitos desempenhados pelo AnoTex responsável pela compilação e anotação do *CorpACE* é demonstrado na Figura 1. Esse conjunto, em síntese, corresponde ao delineamento do modelo proposto na pesquisa, dividido em quatro etapas: *1 seleção*, *2 compilação*, *3 processamento* e *4 exportação* dos dados, a saber:

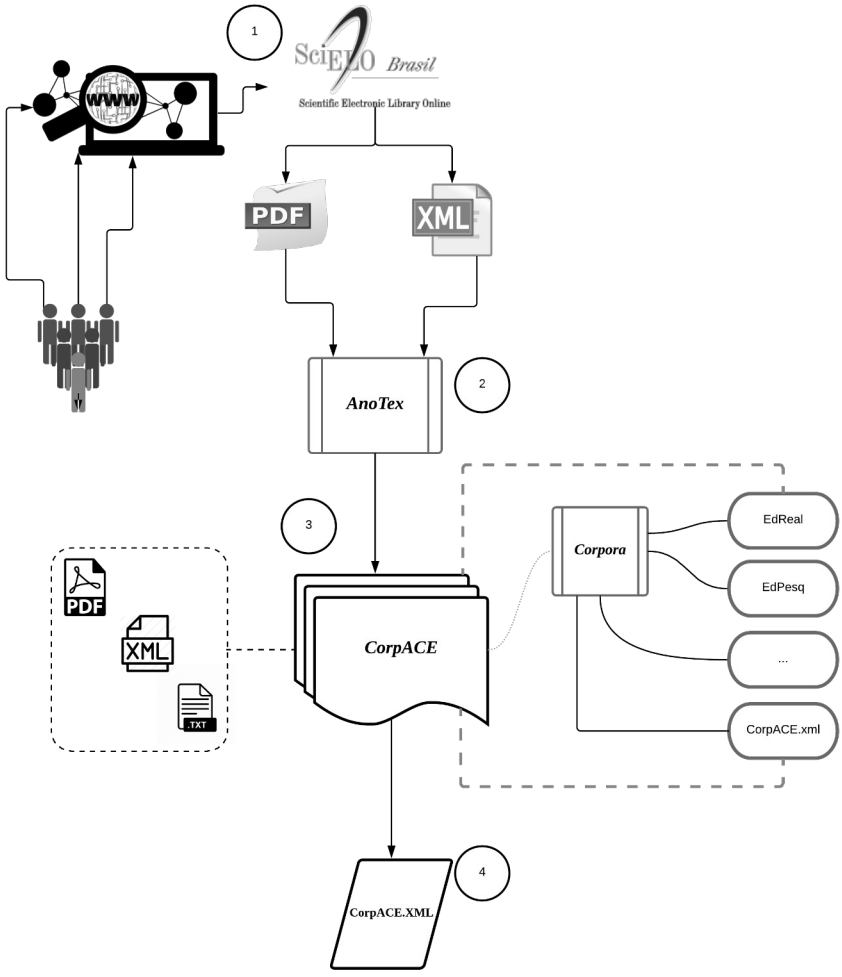

Figura 1: Modelo de compilação do AnoTex.
Fonte: Elaborado pela autora.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

As etapas de *seleção*, *compilação*, *processamento* e *exportação* dos dados desempenhadas pelo modelo computacional em destaque serão detalhadas nas subseções seguintes.

5.1 Etapa 1 - Seleção

A compilação e anotação do *CorpACE* objetivou apontar a configuração de um modelo computacional do gênero artigo científico, enfatizado pelo uso de etiquetas XML, que permitisse destacar as características e a visualização das dimensões constitutivas do gênero e delimitar os objetivos a serem atingidos em relação aos diferentes propósitos de seu uso. Além disso, a representação arbórea dos elementos constitutivos do *corpus* pode dar pistas das características do gênero que podem ser mineradas e valoradas para o processamento do texto.

A escolha dos *corpora,* constituídos por dois *corpus* com oitenta e sete artigos científicos, de domínio educacional, publicados pela Revista Educação & Realidade[18], ISSN: 2175-6236, e pela Revista Educação e Pesquisa[19], ISSN: 1678-4634, ambas *Qualis*[20] A1, disponibilizados no *site* da SciELO, para análise do estudo, foi feita sob a observação dos seguintes critérios de escolha:

1. por sua produção pertencer ao domínio discursivo científico, atividade bastante frequente na esfera acadêmica (função);
2. constituir um tipo relativamente estável de enunciado (estrutura), condicionando sua função social;
3. ser um gênero secundário (modalidade escrita), que possui mais características da escrita do que da fala, por buscar mais objetividade, concisão e padronização;
4. apresentar uma estrutura básica, mínima, dos elementos estruturais do texto, constituída de: título, nome do autor, introdução, corpo do artigo, conclusão e referências;
5. possuir um resumo de referência;
6. ser de domínio público e não necessitar de autorização de uso;
7. estar disponível em formato aberto, em meio *on-line*;
8. possibilitar a filtragem dos dados do XML e do PDF de seu arquivo de origem, estratégia que permite dois módulos de filtragem: do XML, que já é disponibilizado, e o módulo de escaneamento de termos no PDF;
9. permitir a utilização das duas bases de informações para produção de um elemento do *corpus*, cuja representação será uma saída XML.

18 Disponível em: <http://www.seer.ufrgs.br/index.php/educacaoerealidade/index>. Acesso em: 12 jan. 2018.

19 Disponível em: <http://www.educacaoepesquisa.fe.usp.br/>. Acesso em: 12 jan. 2018.

20 O Qualis constitui-se num sistema brasileiro de avaliação de periódicos, mantido pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Relaciona e classifica os veículos utilizados para a divulgação da produção intelectual dos programas de pós-graduação "*stricto sensu*" (mestrado e doutorado), quanto ao âmbito da circulação (local, nacional ou internacional) e à qualidade (A, B, C), por área de avaliação. Os estratos estão divididos em 8 níveis, em ordem de qualidade. (Adaptado de: <https://pt.wikipedia.org/wiki/Qualis>. Acesso em: 27 nov. 2018).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Esse conjunto de requisitos é necessário para impactar na validade e confiabilidade dos *corpora* de análise.

5.2 Etapa 2 – Compilação

Até o momento, a forma de operação do AnoTex é por linha de comando e inspirada em programas de controles de versão (git, svn), em que se tem o comando principal que é a chamada do programa e o subcomando com suas opções, como descrito no exemplo seguinte:

>anotex <subcomando> <opções relevantes ao subcomando>

Na versão AnoTex 0.1 beta linha de comando, estão implementadas as versões experimentais de quatro comandos: “add”, “rm”, “ls” e “stat”. O comando "add" serve para adicionar um novo elemento ao *corpus*. Se o *corpus* não estiver criado é com esse comando que a referida tarefa será executada. E, se esse *corpus* por sua vez precisar estar dentro de uma coleção maior – os *corpora –,* é com esse comando, também, que a referida tarefa será executada. O comando *add* espera por cinco argumentos obrigatórios e tem mais dois opcionais, conforme especificações no Quadro 2, a seguir:

Quadro 2: Linha de comando para adicionar no AnoTex.

	Argumentos obrigatórios para add
--xml --pdf --corpora --corpus	Especifica o caminho para o XML da SciELO relativo ao artigo de referência Especifica o caminho para o PDF do artigo de referência Especifica o nome dos *corpora* onde o *corpus* está inserido ou deverá ser inserido [caso não exista] Especifica o nome do *corpus* que conterá os elementos do artigo de referência
--out	Especifica o caminho para o arquivo XML que representa a coleção
	Argumentos opcionais para add
--txt-dir	Especifica o nome do subdiretório para onde posteriormente uma versão txt desse arquivo será disponibilizada
--txt-file	Especifica o nome de arquivo para esse arquivo txt

Fonte: Elaborado pela autora.

O comando disponível “rm” é usado para remover um elemento do *corpus*, um *corpus* ou um *corpora*. O nível de remoção vai depender do quanto de informação o usuário fornecer para o programa, tudo dependerá dessa especificação. No momento, o AnoTex possui apenas dois argumentos obrigatórios e dois opcionais para remoção, conforme especificado no Quadro 3:

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Quadro 3: Linha de comando para remover do AnoTex.

	Argumentos obrigatórios para rm
--xml	Especifica o caminho para a coleção, na prática sempre será o --out do comando add
--corpora	Especifica o nome dos *corpora* onde a operação de remoção será efetuada
	Argumentos opcionais para rm
--corpus	Especifica o nome do *corpus* onde a operação de remoção será efetuada
--text	Especifica o título do texto para a operação de remoção

Fonte: Elaborado pela autora.

Os comandos “stat” e “ls” possuem finalidades exploratórias, portanto oferecem algumas contagens básicas e possibilitam pesquisar dentro dos *corpora* e seus *corpus*. O *stat* informa o total de *corpora*, lista um por um detalhando os *corpus* e indica quantos textos há dentro de cada um. Além disso, ainda informa a data da última alteração realizada pelos comandos *add* ou *rm*. Já o *ls* lista os textos existentes dentro de um determinado *corpus*. Opcionalmente, com o *ls* é possível explorar um artigo específico, para isso esse comando aceita os *globs* (caracteres curinga). Os *globs* implementados no AnoTex são os mais comuns para pesquisas, como: “*” (corresponde a qualquer número de caracteres, incluindo nenhum), “?” (ao menos uma ocorrência de qualquer caractere), “[]” (corresponde ao caractere dado no colchete). Dessa forma, é possível listar todos os artigos que contenham algum elemento-chave usado na pesquisa.

Caso o usuário necessite de uma sinopse rápida para se lembrar de todos os argumentos de um determinado comando, basta digitar: >anotex help <comando> e as especificações sobre as informações do programa aparecerão na tela. Essa funcionalidade pode ser observada na Figura 2, com as especificações do comando *rm*, a seguir:

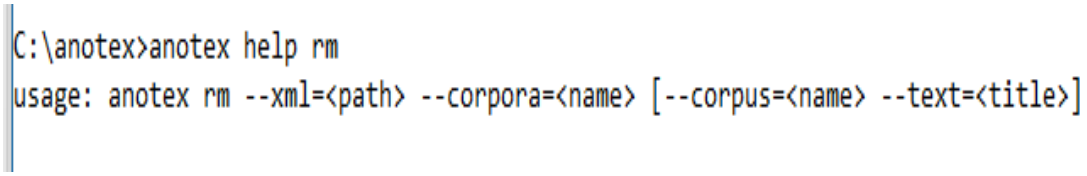

Figura 2: Tela Anotex Ajuda.
Fonte: AnoTex v0.1b.

Além do comando ajuda, o AnoTex também apresenta o comando ←-version>, que exibe a versão do aplicativo, conforme demonstração na Figura 3, a seguir:

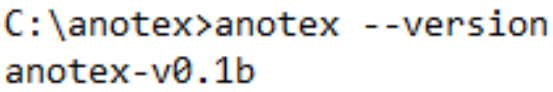

Figura 3: Versão do aplicativo.
Fonte: AnoTex v0.1b

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Até o momento, a versão beta 0.1 dispõe dos comandos “add”, “rm”, “ls” e “stat”. O programa segue a regra do silêncio da filosofia Unix, cuja máxima é “*Rule of Silence: When a program has nothing surprising to say, it should say nothing*.”[21] Então, quando uma operação for bem-sucedida, o programa, provavelmente, vai ficar em silêncio. Quando houver um erro, portanto, ele vai reclamar de algo com alguma mensagem sinalizadora. Outra regra seguida é a da simplicidade: projete para simplicidade; adicione complexidade apenas onde é necessário.

Essa configuração de anotação e compilação do AnoTex permite ao *CorpACE* dar a possibilidade de edição dos textos da forma que o usuário preferir, caso algum dado não tenha sido convertido corretamente. Nesse caso, o usuário pode instrumentar o *corpus* de acordo com suas preferências e propósitos. A Figura 4, a seguir, apresenta uma amostra da configuração do *CorpACE* gerada pelo AnoTex:

```
        <?xml version="1.0" encoding="ISO-8859-1"?>
        <corpora name = "CorpACE" last-change = "1528917394">
                <corpus name = "EdReal" last-change = "1528672710">
                        <text title = "Abordagens do Racismo em Livros DidÃ¡ticos de HistÃ³ria (2008-
2011)" relpath = "C:\anotex/01013.txt">
                                <abstract>O artigo discute as formas pelas quais aspectos relativos
[...]Os Resultados obtidos demonstram que uma diversidade de abordagens tem caracterizado o tratamento
do tema, especialmente, no que tange Ã s finalidades Ã©tico-polÃ-tico-cultural.</abstract>
                <kwd-group>
                        <kwd data = "Ensino de HistÃ³ria"/>
                        <kwd data = "Lei 10.639/03"/>
                        <kwd data = "Atividades"/>
                        <kwd data = "Livros DidÃ¡ticos"/>
                        <kwd data = "PÃ³s-AboliÃ§Ã£o."/>
                </kwd-group>
        <title-group>
                        <article-title data = "Abordagens do Racismo em Livros DidÃ¡ticos de HistÃ³ria
(2008-2011)"/>
                </title-group>
                        <sections>
                        <section title = "IntroduÃ§Ã£o"/>
                        <section title = "Apontamentos acerca da Historiografia sobre o PÃ³s-aboliÃ§Ã£o
e suas RepercussÃµes DidÃ¡ticas"/>
                        <section title = "ReflexÃµes sobre Atividades na HistÃ³ria Escolar e em Livros
DidÃ¡ticos"/>
                        [...]
                        <section title = "ConsideraÃ§Ãµes Finais"/>
                        </sections>
                        <references>
                                <ref-entry>
                                        [...]
                                </ref-entry>
                                </references>
                                <markups>
                                        [...]
                                        <markup data = "recuperar" freq = "1" page = "15" />
                                        <markup data = "imprensa negra" freq = "1" page = "15" />
                                        <markup data = "organizações negras" freq = "1" page = "15" />
                                        <markup data = "referências " freq = "1" page = "15" />
                                        <markup data = " " freq = "1" page = "15" />
                                        <markup data = " topoi " freq = "1" page = "15" />
                                        [...]
                                </markups>
                        </text>
        [...]
                </corpus>
        </corpora>
```

Figura 4: Amostra da configuração do CorpACE.
Fonte: AnoTex v0.1b.

21 Regra de silêncio: quando um programa não tem nada de surpreendente em dizer, não deve dizer nada. (tradução nossa) <http://www.linfo.org/rule_of_silence.html>.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

5.3 Etapa 3 – Processamento

O AnoTex possui programa multi-plataforma e utiliza uma rotina que filtra automaticamente essas informações que, quando carregadas em memória, respeitam a hierarquia ditada no XML. As informações disponibilizadas em XML no *site* da SciELO podem ser exportadas e filtradas por uma rotina baseada nas especificações do Guia de uso de elementos e atributos XML para documentos, que seguem a implementação *SciELO Publishing Schema*, Versão 1.5.1 – setembro de 2016[22]. A análise dos elementos constantes da seção “Lista de Elementos”, compreendidos em mais de uma centena de marcações, desde “<abbrev-journal-title>” até “<year>”, podem ser suportados pela ferramenta. Dessa forma, a rotina encontrando alguma informação marcada com essas etiquetas XML, compreendidas nesse intervalo, é capaz de filtrá-la para memória do *CorpACE* para uma busca posterior.

Ao final do processamento do arquivo XML, tem-se uma estrutura em árvore na memória do computador seguindo a hierarquia imposta pelo próprio XML. Buscas poderão ser feitas, informações poderão ser acessadas e exportadas conforme o interesse do anotador. O limite de informação seria imposto no próprio XML disponível no *site* da SciELO pelas revistas e periódicos.

Para a filtragem e exportação dos conteúdos em PDF para HTML, é utilizada uma ferramenta gratuita que captura qualquer sequência de elemento de destaque em negrito e em itálico ao longo do artigo. A referida ferramenta é o “pdftohtml[23]” (programa para converter arquivos pdf em imagens html, xml e png), que vem por padrão em quase toda distribuição Linux para *desktop* e há também uma versão para Windows. Essa ferramenta gera sua saída no atual diretório de trabalho. Sendo o PDF um formato muito complexo, bem parecido com um arquivo de imagem, nem sempre a conversão para HTML é ótima. Em função da dificuldade de filtrar as marcações em negrito e itálico do formato PDF, pode-se usar compressão de dados, uma vez que nesse caso é difícil *parsear* o arquivo em busca dessas marcações. Com isso, é mais fácil utilizar uma ferramenta que converte a entrada PDF para uma saída mais textual e marcada que é o HTML. O HTML por sua vez possui as *tags* <B>...</B> para negrito e <I>...</I> para itálico. Então, tiramos vantagem do HTML e do pdftohtml para vencer a complexidade imposta pelo formato (caixa-preta) PDF. Alguns dos elementos filtrados podem necessitar de uma edição posterior por um anotador humano ou uma correção por outras ferramentas. A intenção, nesse momento da pesquisa, é filtrar apenas os conteúdos que o pdftohtml conseguir marcar e combiná-los com os elementos filtrados do XML do Guia SciELO. Por meio dessa filtragem, também é possível saber quantas vezes e em qual página (numeração) o termo marcado ocorreu. Essa informação pode se tornar relevante, uma vez que os elementos mais conclusivos e significativos podem aparecer em páginas mais próximas da seção introdução ou conclusão nos artigos científicos.

22 Disponível em: <http://docs.scielo.org/projects/scielo-publishing-schema/pt_BR/1.5-branch/>. Acesso em: 10 dez. 2017.

23 Pdftohtml foi desenvolvido por Gueorgui Ovtcharov e Rainer Dorsch. Baseia-se e beneficia muito do pacote xpdf da Derek Noonburg. Disponível em: <http://www.tutorialspoint.com/unix_commands/pdftohtml.htm>. Acesso em: 18 jan. 2018.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

5.4 Etapa 4 – Exportação

Em relação à anotação, segundo Aluísio e Almeida (2006), são basicamente dois os níveis de representação das informações dos elementos estruturais do artigo científico que são representadas no *corpus*: a anotação estrutural e a anotação linguística.

A *anotação estrutural* compreende a marcação de dados externos e internos dos artigos científicos. Como dados externos, entendemos a documentação do *corpus* na forma de um cabeçalho que inclui os metadados dos elementos pré-textuais. Como dados internos temos a anotação de segmentação dos elementos textuais e pós-textuais.

A *anotação linguística* pode ser em qualquer nível que se queira, isto é, nos níveis morfossintático, sintático, semântico, discursivo, fonológico, etc., sendo inserida de três formas: manualmente (por linguistas), automaticamente (por ferramentas de PLN) ou semiautomaticamente (correção manual da saída de outras ferramentas).

O padrão usado na anotação do *CorpACE* seguiu a marcação estrutural disponível pelo banco de dados SciELO. A segmentação dos elementos a serem anotados seguiu a arquitetura imposta pelo arquivo XML de origem do artigo, conforme demonstrado na Figura 5, a seguir:

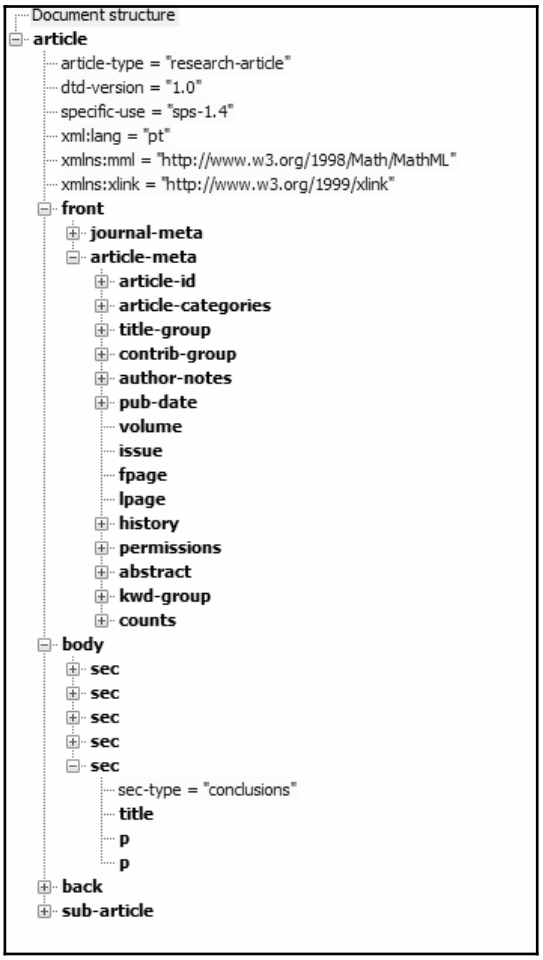

Figura 5: Demonstração em árvore do arquivo XML do artigo científico.
Fonte: Banco de dados da SciELO – Adaptado pelos autores.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

Nessa etapa, o AnoTex cria o *CorpACE* gerando um arquivo de saída XML (baseado na entrada dos dados do XML e do PDF). Essa base textual servirá para análise linguística, treinamento e análise de ferramentas computacionais cuja finalidade seja o processamento de texto.

Para a saída dos *corpora* em um arquivo XML, o AnoTex precisou receber os 05 argumentos obrigatórios: o arquivo em PDF, o arquivo XML, a denominação dos *corpora*, a denominação do arquivo de saída e em qual *corpus,* desse arquivo de saída, a marcação foi adicionada. Essa opção *corpus* é importante, pois se o arquivo de saída estiver vazio, ou não existir, o *corpus* será inicializado com a nova marcação. Entretanto, se o *corpus* já estiver constituído, a nova marcação será adicionada no final do mesmo. Uma vantagem, desse processo, é que um arquivo XML de saída pode reunir diversos *corpus*, requisito muito útil se, por exemplo, for necessária a criação de diferentes versões do mesmo *corpus*. O AnoTex permite a criação de vários *corpus,* adicionados dentro do mesmo arquivo, esse conjunto de requisito constitui os *corpora* do arquivo. Todas essas funcionalidades puderam ser visualizadas nas Figuras 1 e 4.

6 Discussão

Dentre os elementos que podem ser filtrados pelo AnoTex, destacam-se a estrutura do gênero do texto-base, a pontuação relacionada à frequência de termos em todo o documento, o título, os subtítulos, expressões destacadas – como negrito e itálico – localizadas na sentença e a ocorrência de informações não essenciais. Todas essas características que são filtradas, compiladas e exportadas pela ferramenta computacional vão gerar um arquivo de saída em formato XML. Uma das vantagens desse tipo de configuração de arquivo de saída é permitir a criação de um *corpus* que pode ser usado por um número maior de ferramentas e de forma que atenda a propósitos variados, seguindo via de regra, os dados que são disponibilizados pelo arquivo XML SciELO.

A reutilização do *CorpACE* por outras ferramentas torna o trabalho mais relevante, motivo pelo qual as configurações são deixadas mais abertas e genéricas, para abrir possibilidades para novas estratégias e assim não engessar tanto a estrutura do *corpus*, deixando-o mais maleável aos novos propósitos para uso. Outro ponto importante a ser destacado é que cada elemento que compõe o *corpus* inclui um caminho relativo (*relpath*), nesse caso o *corpus* seria composto do arquivo XML principal (‘CorpACE.xml’) e de uma subpasta (‘texts’) contendo todos os textos na íntegra em txt. Esse caminho é importante, pois quando uma ferramenta para processamento de texto lesse uma seção <text>...</text>, saberia exatamente onde estaria o texto correspondente na íntegra, utilizando esse procedimento.

As marcações em XML das representações de aspectos relacionados com o **contexto de produção** do artigo científico podem influenciar na forma como o texto se organiza. Os elementos filtrados constitutivos dessas marcações revelam: emissor/enunciador, receptor/destinatário, lugar/instituição, quando/tempo e espaço, suporte, conteúdo temático. Já a **arquitetura geral** do gênero artigo que está relacionada ao gerenciamento do conteúdo do texto – a construção composicional característica do gênero –, a princípio, são filtradas pelo AnoTex as capacidades discursivas do gênero

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

compreendidas em três subdivisões:

1. Elementos **Pré-textuais**, como: título, autor, instituição, resumo e palavras-chave;
2. Elementos **textuais**, como: introdução, discussão e considerações finais (principais seções do artigo);
3. Elementos **Pós-textuais**, como: referências bibliográficas.

Essa configuração de modelo computacional do gênero artigo científico, enfatizado no uso de etiquetas XML, permite destacar as características e a visualização das dimensões constitutivas do gênero, e delimitar os objetivos a serem atingidos em relação aos diferentes propósitos de seu uso. Além disso, essa representação arbórea dos elementos constitutivos do *corpus* pode dar pistas das características do gênero que podem ser mineradas e valoradas para o processamento do texto.

A indicação da localização e quantidade de ocorrências da filtragem dos elementos em negrito e itálico podem se tornar relevantes, uma vez que os elementos mais conclusivos e significativos podem aparecer em páginas mais próximas da seção introdução ou conclusão nos artigos científicos. Nesse caso, poderia ser estabelecido um peso para o termo, observando-se a página onde ele ocorresse, além do número de vezes que ele fosse encontrado e marcado ao longo de todo artigo. Do ponto de vista linguístico, essa é uma informação que ainda precisa ser explorada na valoração do ranqueamento das sentenças.

7 Conclusão

As ferramentas de PLN pretendem simular as principais características humanas para identificar segmentos relevantes de um ou mais textos-fonte e transformá-los em um novo texto, cuja simulação implica a prática da retextualização automática. A descrição do gênero como objeto linguístico, concreto via artigo científico, revelou-se um importante recurso para análises de processamento automático de texto. Sua importância se dá pela análise, seguindo diferentes pontos de vista teóricos (computacional e linguístico), que podem evidenciar elementos constitutivos do objeto, tais como a recuperação da estrutura básica do gênero artigo científico constituída de elementos pré-textuais (<front>), textuais (<body>) e pós-textuais (<back>). Essas evidências permitem partir para explicações e, posteriormente, para generalizações e previsibilidades que um *corpus* pode revelar. Feito esse percurso, pode-se retornar ao ponto de partida teórico e verificar o quanto o objeto, efetivamente, pode dar suporte como técnica que está ao alcance da tecnologia em PLN e que está à disposição do pesquisador no momento, mesmo relativamente rasa.

Uma vez que o AnoTex faz uso de XML para anotação da estrutura do gênero textual, evidenciou-se a vantagem de reutilizar a mesma tecnologia por outros anotadores (eDictor, COMEDI, Aelius) para anotações morfológica e sintática de edições de *corpora*. Como o XML é um padrão, usá-lo para todas as representações nos textos do *corpus* favorece a criação de recursos padronizados, permitindo reuso de tecnologia, oferecendo mais flexibilidade para as buscas e exibição dos resultados e independência tecnológica para grupos de pesquisa interessados em estudo nesse *corpus*.

Como foi evidenciado, os editores, atualmente disponíveis, têm as suas forças individuais em se tratando da cobertura da anotação morfossintática, sintática e

Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

semântica para *corpora*. Entretanto, também apresentam deficiências, em particular, no que se refere à anotação dos elementos constitutivos da estrutura básica do gênero textual, que também constitui relevante anotação linguística para o processamento do texto base. O AnoTex oferece facilidade de utilização e traz uma especificação para anotação dos elementos constitutivos do gênero textual, como recurso de linguagem em XML, que pode ser aproveitado no PLN. Sendo assim, presumimos que nosso trabalho em PLN preenche uma lacuna atual e esperamos que seja útil aos pesquisadores em toda a infraestrutura textual com a anotação do gênero textual de artigos científicos, adotado pelo AnoTex.

Referências

ALENCAR, L. F. *CORPTEXLIT – Corpus* de Língua Portuguesa de Textos Literários do Século XIX. Fortaleza: [s.n.], 2010. Disponível em: <http://complin.blogspot.com.br/2012/03/corpus-de-textos-historicos.html>. Acesso em: 19 abr. 2018.

ALENCAR, L. F. Novos recursos do Aelius para o processamento computacional raso do português. *Dialogar é preciso*: linguística para o processamento de línguas. Vitória: PPGEL/UFES, 2013.

ALENCAR, L. F. *About Aelius Brazilian Portuguese POS-Tagger*. Brasil, 2013a. Disponível em: <http://aelius.sourceforge.net/>. Acesso em: 19 abr. 2018.

ALENCAR, L. F. *Aelius User's Manual*. UFC, 2013b. Disponível em: <http://aelius.sourceforge.net/manual.html>. Acesso em: 19 abr. 2018.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, v. 4, n. 3, p. 156-178, 2006.

BAKHTIN, M. M. *Estética da criação verbal*. São Paulo: Livraria Martins Fontes, 1992.

BHARTI, S. K.; BABU, K. S. *Automatic Keyword Extraction for Text Summarization*: A Survey. arXiv preprint arXiv:1704.03242, 2017.

CAMBRIA, E.; WHITE, B. Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, v. 9, n. 2, p. 48-57, 2014

CANALES, C. B. La red SciELO (Scientific Electronic Library Online): perspectiva tras 20 años de funcionamiento. *Hospital a Domicilio*, v. 1, n. 4, p. 211-220, 2017.

COHEN, J. D. Highlights: language-and domain-independent automatic indexing terms for abstracting. *Journal of the American society for information science*, v. 46, n. 3, p. 162-174, 1995.

Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

DA COSTA CARVALHO, C. I.; VASCONCELOS, D. M.; DE ALENCAR, L. F. Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem. *Anais do X Encontro de Linguística de Corpus*: Aspectos metodológicos dos estudos de corpora. Belo Horizonte: Editora da UFMG, p. 122-134, 2012.

DA SILVA, B. C. D. O estudo lingüístico-computacional da linguagem. *Letras de Hoje*, v. 41, n. 2, p. 103-138, 2006.

DIMA, E. et al. A Metadata Editor to Support the Description of Linguistic Resources. In: *LREC,* 2012. p. 1061-1066.

EDICTOR – Humanidades Digitais. Grupo de Pesquisas da Universidade de São Paulo. Brasil. Disponível em: <https://humanidadesdigitais.org/edictor/>. Acesso em: 19 abr. 2018.

FARIA, P.; GALVES, C. Criando “bancos de árvores”: o sistema de anotação e o processamento automático. *Cadernos de Estudos Linguísticos*, v. 58, n. 2, p. 299-315 2016.

FIALHO, P. et al. Inesc-id@ assin: Medição de similaridade semântica e reconhecimento de inferência textual. *Linguamática*, v. 8, n. 2, p. 33-42, 2016.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, v. 47, n. 1, p. 1-66, 2017.

GUIA de uso de elementos e atributos XML para documentos que seguem a implementação SciELO Publishing Schema. Versão 1.5.1 – setembro de 2016. Disponível em: <http://docs.scielo.org/projects/scielo-publishing-schema/pt_BR/1.5-branch/>. Acesso em: 01 ago. 2018.

HOVY, E.; NAVIGLI, R.; PONZETTO, S. P. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, v. 194, p. 2-27, 2013.

KING, M. et al. Variability of the Facet Values in the VLO–a Case for Metadata Curation. In: *Selected Papers from the CLARIN Annual Conference 2015*, October 14–16, 2015, Wroclaw, Poland. Linköping University Electronic Press, 2016. p. 25-44.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, v. 25, n. 2-3, p. 259-284, 1998.

LOVINS, J. B. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, v. 11, n. 1-2, p. 22-31, 1968.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of research and development*, v. 2, n. 2, p. 159-165, 1958.

LYSE, G. I.; MEURER, P.; DE SMEDT, K. COMEDI: A component metadata editor. In:

Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands. Linköping University Electronic Press, 2015. p. 82-98.

MANARIS, B. Natural language processing: A human-computer interaction perspective. *Advances in Computers*, v. 47, p. 1-66, 1998.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. In: DIONÍSIO, A. P; MACHADO, A. R; BEZERRA, M.A (org). *Gêneros textuais e ensino*. 4ª ed. Rio de Janeiro: Lucerna, 2005, p. 19-36.

MATENCIO, M. de L. M. Atividade de (Re) textualização em práticas acadêmicas: um estudo do resumo. *Scripta*, v. 6, n. 11, p. 109-122, 2002.

PACKER, A. L. SciELO: uma metodologia para publicação eletrônica. *Ciência da informação*, v. 27, n. 2, p. 109-121, 1998.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. de. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. *Caminhos da linguística de corpus*. Campinas: Mercado de Letras, 2010, p. 191-224.

PAIXÃO DE SOUSA, M. C. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, v. 16, n. spe, p. 53-93, 2014.

REITER, E. A Structured Review of the Validity of BLEU. *Computational Linguistics*, n. Just Accepted, p. 1-12, 2018.

ROCHA, V. C.; GUELPELI, M. V. C. “PragmaSUM: automatic tex summarizer based on user profile”, *International Journal of Current Research*, Vol. 9, Issue, 07, p. 53935-53942, July, 2017.

SANTOS, C. D.; ZADROZNY, B. Learning character-level representations for part-of-speech tagging. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14),* 2014. p. 1818-1826.

SANTOS, S. de J. B. dos. A importância da leitura no ensino superior. *Revista de educação*, v. 9, n. 9, p. 77-83, 2015.

SOUZA, L. F. C. de; FARIA, P. P. F. De; TEMPONI, C. N. Uma proposta de automatização das edições XML do e-Dictor. VIII SEMINÁRIO DE ESTUDOS FILOLÓGICOS–SEF, FILOLOGIA E HUMANIDADES DIGITAIS, 2016. *Anais…* 2016. Disponível em: <https://sefuefs2015.wordpress.com/uma-proposta-de-automatizacao-das-edicoes-xml-do-e-dictor/>. Acesso em: 01 ago. 2018.

VIEIRA, R.; LIMA, V. L. S. de. Lingüística computacional: princípios e aplicações. In: *Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial*. sn, 2001. p. 47-86.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 40-64, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.40-64

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in NLP. In: *Proceedings of the 14th conference on Computational linguistics-Volume 4*. Association for Computational Linguistics, 1992. p. 1106-1110.

Recebido em dia 05 de julho de 2018.
Aprovado em dia 03 de outubro de 2018.