ANÁLISE DE TEXTOS E CRIAÇÃO AUTOMÁTICA DE ATIVIDADES DE LEITURA EM INGLÊS COM CORPORA

Text analysis and automatic creation of English reading activities using *corpora*

José Lopes MOREIRA FILHO (Secretaria de Educação do Estado de São Paulo, Brasil)

RESUMO

O uso de corpora no ensino de línguas é tema recorrente em Linguística de Corpus. O processo de criação de atividades para o ensino de línguas pode ser amplamente beneficiado por meio do uso de dados da exploração de corpora e ferramentas computacionais de análise linguística. Este estudo descreve um sistema de análise de textos e criação automática de atividades de leitura em língua inglesa. Nota-se que os resultados potencializam a garantia de materiais de ensino que privilegiam a língua em uso e também fornecem análises linguísticas variadas, com menor esforço humano, para a tarefa de elaboração de atividades didáticas.

Palavras-chave: *Linguística de Corpus; Ferramentas computacionais; Leitura; Ensino de Línguas*.

ABSTRACT

The use of corpora in language teaching is an important topic, since practice is aimed at ensuring that the teaching material is focused on the language in use. The process of creating activities for language teaching can be improved by using corpus data and computational tools in linguistic analysis. This study describes a system for text analysis and automatic creation of English reading activities. The results show that the system allows the development of teaching materials that focus on language in use and it also provides varied linguistic analysis, with less human effort, to the task of developing reading activities.

Key-words: *Corpus Linguistics; Computational tools; Reading; Language Teaching.*

1. Introdução

Considera-se que o uso de materiais baseados em *corpus* é desejável para o processo de ensino-aprendizagem, mas a sua preparação e elaboração ainda não é uma realidade comum, principalmente por professores fora do contexto acadêmico. As ferramentas computacionais disponíveis e o conhecimento específico na área da Linguística de Corpus parecem ainda estar restritos às pesquisas científicas.

A tarefa de elaboração de materiais didáticos baseados em *corpora*, mesmo por pesquisadores, demanda tempo e esforço; muitas vezes, requer a análise prévia de grandes quantidades de dados por programas de computador especializados, como concordâncias, listas de frequência, listas de palavras-chave, anotação de *corpus*, entre outros tipos de análise. Muitas das ferramentas são de uso geral para a análise de *corpora* e suas diversas aplicações, não sendo específicas para o ensino de línguas ou para a tarefa de elaboração de material didático.

Tendo em vista a problemática, professores podem ter dificuldades na preparação de tais materiais e, em consequência, não utilizá-los. Verifica-se também uma lacuna em relação a ferramentas computacionais voltadas para a elaboração de atividades baseadas em análise de *corpora*. Mesmo os pesquisadores e especialistas da área, em tal cenário, têm dificuldades em encontrar programas que atendam a necessidades específicas. É comum, por exemplo, o uso de planilhas do *Microsoft Excel* para filtrar listas de itens lexicais e realizar anotações adicionais aos resultados de análise de listadores de palavras e concordanciadores.

Desse modo, acredita-se que a preparação de materiais no ensino de línguas pode ser assistida e enriquecida pelo auxílio de recursos computacionais projetados especificamente para tal fim, os quais possam contribuir para a otimização dos estudos linguísticos em larga escala, tanto em relação a tempo e esforço, quanto em qualidade de seus resultados, por meio da automatização de análises de textos na exploração de *corpora* e, também, do aproveitamento dos dados de análise para a criação de atividades.

A partir do desenvolvimento de um sistema de criação automática de atividades de leitura em língua inglesa com *corpora*, tenta-se suprir a necessidade de professores de língua estrangeira que desejam preparar e utilizar materiais baseados em *corpora*, embora não estejam familiarizados com o uso de ferramentas de processamento e exploração de *corpora* e/ou que não possuem muito tempo para todo o processo de preparação.

Para tanto, com base em abordagens e ferramentas de análise de *corpora* da área de Linguística de Corpus, bem como em técnicas, algoritmos e recursos de áreas da Computação, como o Processamento de Línguas Naturais e o Aprendizado de Máquina, é proposta[1] a criação de um Sistema de Processamento de Língua Natural capaz de analisar textos para gerar, automaticamente, atividades de leitura e ensino de padrões (léxico-gramática) em língua inglesa a partir de um texto e *corpus*.

2. Ensino de leitura e criação de atividades

A proposta de um sistema para criação automática de atividades de leitura em língua inglesa relaciona-se a aportes teóricos e metodológicos para o ensino de leitura, além de análises de materiais de ensino de leitura instrumental do contexto brasileiro. A revisão dos estudos contribuiu para a instanciação de possibilidades e a construção de uma proposta mais consistente no desenvolvimento do sistema em relação a uma série de questões técnicas e pedagógicas: estágios do processo de criação de atividades, elaboração de unidades didáticas, procedimentos de ensino de leitura com base em gêneros textuais, tipos de atividades e sua aplicabilidade a diferentes textos. As considerações sobre os estudos que influenciaram a construção do sistema, tecidas brevemente aqui, são importantes para o entendimento da natureza de tais questões.

Um dos aportes de base é o uso do conceito de *standard exercise*[2] (SCOTT *et al.*, 1984), em que um rol de perguntas na língua do aprendiz pode ser aplicado a quase qualquer texto, a fim de treinar estratégias de leitura em língua inglesa, de forma a facilitar a preparação de materiais novos e maximizar as opções de escolha de atividades com texto por aprendizes. A possibilidade de elaboração de atividades padrão para textos valida a ideia de extrapolação desse procedimento, no estudo, a partir da criação de modelos de atividades reutilizáveis que podem ser aplicados a diferentes textos a partir de análises automáticas.

Sobre como estruturar atividades de leitura, Grellet (1981) apresenta um conjunto de itens para a confecção de uma aula de leitura baseada em texto: utilização de estratégias; interpretação do texto; algum tipo de produção que considere o texto como um todo. Holmes (1982) fornece um guia para a preparação de materiais para leitura em um curso de inglês

[1] Tal proposta refere-se a uma pesquisa de doutorado (MOREIRA FILHO, 2015).
[2] Atividade padrão, tradução nossa.

para fins específicos, o qual também aborda questões de estratégias de leitura e atividades, a partir de níveis de compreensão do texto, estratégias para facilitar a leitura e tipos de exercícios.

Em relação a leitura e gêneros textuais, Ramos (2004) considera que o trabalho pode auxiliar o professor tanto em seu melhor entendimento do texto (enquanto gênero), quanto em relação ao que os alunos têm de fazer linguisticamente, e propõe uma prática de ensino a partir de três fases: apresentação, detalhamento e aplicação.

Em análise a livros didáticos para o ensino de leitura de inglês, em abordagem instrumental, determinados itens de ensino são recorrentes e podem ser utilizados como referência para ampliação das possibilidades de criação automática de exercícios, conforme permite o conjunto de análises e instrumentação computacional disponível: reconhecimento de gêneros textuais, níveis de compreensão de leitura, estratégias de leitura e aspectos léxico-gramaticais.

O estudo das diversas propostas para a preparação de atividades de leitura em língua estrangeira e tipos de exercícios teve como objetivo explorar as possibilidades de criação de atividades a partir de análises automáticas de texto e *corpora*, assim como buscar quadros conceituais flexíveis e adaptáveis para o processo de elaboração de atividades para o ensino de leitura, a fim de colaborar para o desenvolvimento do sistema.

3. Linguística de Corpus: ferramentas e procedimentos básicos

A Linguística de Corpus pode ser considerada uma área que estuda a língua por meio da observação de grandes quantidades de dados linguísticos reais por meio do uso de ferramentas computacionais (BERBER SARDINHA, 2004, p. 3). O desenvolvimento da Linguística de Corpus tem relação estreita com os avanços da tecnologia, especificamente o advento do computador, que tem permitido o estudo empírico dos fenômenos da linguagem em números de textos cada vez maiores, com velocidade, precisão e confiabilidade estatística. Há uma série de ferramentas computacionais disponíveis para pesquisa na área. A maioria delas executa tarefas de contagem de itens, cálculos estatísticos, comparações e organização de resultados em visualizações privilegiadas, a fim de que padrões linguísticos sejam identificados com maior facilidade pelo linguista.

Dentre as ferramentas mais utilizadas em Linguística de Corpus estão os programas listadores de palavras e concordanciadores, que podem ser encontrados em uma única solução *desktop*, para análise de *corpora* em geral. Há também ferramentas *online* do mesmo tipo atreladas a *corpora* específicos que, muitas vezes, não permitem o uso de textos fornecidos por usuários.

Os programas listadores de palavras, realizam a contagem de itens em um *corpus* e disponibilizam os resultados em listas de frequência, que podem ser ordenadas conforme a necessidade. A partir das listas, é possível descobrir quais são os itens mais utilizados em um texto ou conjunto de textos. As descobertas podem auxiliar na escolha de itens lexicais para a confecção de atividades e/ou definição de currículos, por exemplo, no contexto do ensino de línguas.

Os concordanciadores são programas utilizados para a extração e a exibição de linhas de concordância. O termo concordância refere-se à listagem das ocorrências de uma palavra de busca de um *corpus*, a qual fica centralizada, com uma quantidade definida de contextos em ambos os lados (esquerda e direita), que permite a descoberta de padrões léxico-gramaticais de uma amostra estudada. Os resultados de análise e a própria visualização privilegiada utilizada podem figurar como matéria prima para a elaboração de atividades de ensino.

As ferramentas descritas anteriormente podem ser aplicadas para os seguintes passos em pesquisas, compondo um conjunto de procedimentos básicos, em nível menos profundo, para a elaboração de cursos e extração de material para confecção de atividades pedagógicas (MOREIRA FILHO, 2015):

a. Coleta do *corpus* – manual, ou automática a partir de programas específicos;
b. Preparação do *corpus* para as análises;
c. Limpeza e organização dos textos;
d. Análise das frequências das palavras do *corpus*;
e. Busca de indícios de palavras que se destacam no *corpus*.
f. Análise de palavras-chave do *corpus*;
g. Busca de palavras que se destacam no *corpus*;
h. Delimitação de um número de palavras para possível análise;
i. Análise da padronização das palavras selecionadas por meio de concordâncias, listas de colocados e n-gramas;

j. Descrição da padronização das palavras selecionadas como resultado;
k. Utilização dos padrões em objetivos seguintes (confecção de atividades, por exemplo).

Os passos podem ser cíclicos e envolver algum tipo de recurso adicional durante o processo de análise e exploração de *corpus*. A descrição e a análise dos procedimentos de exploração de *corpora* podem alçar oportunidades de automatização de partes do processo para o seu aproveitamento em um sistema funcional direcionado a elaboração de atividade a partir de texto e *corpora*.

Na pesquisa, o conhecimento do funcionamento das principais ferramentas e metodologias utilizadas na análise e exploração de *corpora* foi utilizado, com métodos e ferramentas de trabalho das áreas do Processamento de Línguas Naturais e Aprendizado de Máquina, na programação de módulos de análise de texto e *corpora* para o desenvolvimento do sistema.

4. Desenvolvimento do sistema

Para o desenvolvimento do sistema, utilizou-se a linguagem de programação *Python*, com a biblioteca *Natural Language Toolkit* (*NLTK*) (BIRD et. al., 2009). A linguagem foi utilizada em toda a construção dos módulos de análise de texto e *corpora*. O *NLTK* foi utilizado para a criação de etiquetadores e classificadores para as análises linguísticas, além de funções auxiliares e recursos para a avaliação dos algoritmos implementados. O *NLTK* também forneceu recursos linguísticos como *corpora* de referência.

Os *corpora* utilizados na pesquisa foram: o *corpus* de referência *British National Corpus* (*BNC*), o *corpus* de referência Floresta Sinta(c)tica do *NLTK*, o *corpus* de referência MacMorpho do *NLTK*, o *corpus* de estudo/treinamento com 135 textos de anúncios de emprego em inglês da Internet, o *corpus* de estudo/treinamento com 771 textos de divulgação científica das revistas eletrônicas *Scientific American* e *NewScientist*.

O *corpus BNC* foi utilizado na complementação dos dados do etiquetador morfossintático treinado e para gerar uma lista de frequência de palavras que possibilitou a criação do módulo de extração de palavras-chave. Os *corpora* Floresta Sinta(c)tica e MacMorpho foram utilizados para o desenvolvimento do módulo de identificação de palavras cognatas - os radicais das formas em língua portuguesa foram extraídos a fim de serem

comparados com os radicais de palavras em língua inglesa para determinar candidatos a palavras cognatas. Os *corpora* de estudo serviram para o treinamento, desenvolvimento e testes dos principais módulos de análise, além de oferecerem recursos linguísticos para a criação de atividades.

De maneira geral, o processo de desenvolvimento do sistema envolveu a criação de módulos de análise de texto e *corpora*, módulos para a leitura e extração de análises linguísticas, além de módulos para criação de atividades.

5. O sistema de criação automática de atividades de leitura

O sistema desenvolvido é composto por módulos que realizam a análise linguística de um texto de entrada, extraem informações do texto analisado em formato XML por meio de scripts de uma metalinguagem criada para montagem de atividades a partir de informações extraídas dos textos analisados.

Os módulos de análise realizam as tarefas de: segmentação do texto em parágrafos e sentenças; itemização das palavras nas sentenças; extração de raízes das palavras; etiquetagem morfossintática; contagem de frequência das palavras; extração de palavras-chave; identificação de faixas de frequência lexical; identificação de palavras cognatas; identificação de grupos nominais; identificação de grupos verbais; identificação de entidades nomeadas e identificação de afixos.

As análises realizadas pelos módulos são armazenadas em uma estrutura de dados específica, por meio da linguagem XML (*eXtensible Markup Language*), a partir de *tags* personalizadas, a fim de serem acessadas e manipuladas facilmente para a extração de informações e possibilitar análises adicionais. O Quadro 1 exemplifica a estrutura geral em XML utilizada:

Quadro 1 - Estrutura de representação em XML para armazenar as análises

```
<xmltext>
        <genre>
                <category></category>
        </genre>
        <statistics>
                <types></types>
                <tokens></tokens>
                <typetokenratio></typetokenratio>
                <sentences></sentences>
                <paragraphs></paragraphs>
                <cognates></cognates>
        </statistics>
        <title></title>
        <text>
                <paragraph>
                        <sentence>
                                <token></token>
                        </sentence>
                </paragraph>
        </text>
</xmltext>
```

Fonte: Produção do próprio autor.

Durante o processamento, os resultados das análises são utilizados para o preenchimento da estrutura em XML. A Figura 1 ilustra um exemplo de análise realizada e o preenchimento da estrutura no nível de uma sentença:

Figura 1 - Exemplo de estrutura preenchida com dados de análises.

```
<sentence id="7">
    <token cog="0" freq="50" grp="I-NP" id="247" key="4" loc="33.07" pos="DT" size="3" stem="The" suf="-">The</token>
    <token cog="1" freq="3" grp="I-NP" id="248" key="78" loc="33.2" pos="NN" size="7" stem="torrent" suf="-">torrent</token>
    <token cog="0" freq="6" grp="I-NP" id="249" key="85" loc="33.33" pos="NN" size="4" stem="file" suf="-">file</token>
    <token cog="0" freq="3" grp="I-VG" id="250" key="3" loc="33.47" pos="MD" size="4" stem="will" suf="-">will</token>
    <token cog="0" freq="1" grp="I-VG" id="251" key="8" loc="33.6" pos="VB" size="4" stem="tell" suf="-">tell</token>
    <token cog="0" freq="1" grp="I-VG" id="252" key="12" loc="33.73" pos="VB" size="6" stem="anyon" suf="-">anyone</token>
    <token cog="1" freq="1" grp="I-NP" id="253" key="15" loc="33.87" pos="JJ" size="10" stem="interest" suf="ed">interested</token>
    <token cog="0" freq="8" grp="O" id="254" key="2" loc="34.0" pos="IN" size="2" stem="in" suf="-">in</token>
    <token cog="0" freq="1" grp="I-VG" id="255" key="52" loc="34.14" pos="VBG" size="11" stem="download" suf="ing">downloading</token>
    <token cog="0" freq="6" grp="I-NP" id="256" key="6" loc="34.27" pos="DT" size="4" stem="thi" suf="-">this</token>
    <token cog="0" freq="4" grp="I-NP" id="257" key="56" loc="34.4" pos="NN" size="7" stem="content" suf="-">content</token>
    <token cog="0" freq="1" grp="O" id="258" key="3" loc="34.54" pos="WRB" size="3" stem="how" suf="-">how</token>
    <token cog="0" freq="17" grp="I-VG" id="259" key="0" loc="34.67" pos="TO" size="2" stem="to" suf="-">to</token>
    <token cog="1" freq="1" grp="I-VG" id="260" key="14" loc="34.81" pos="VB" size="7" stem="contact" suf="-">contact</token>
    <token cog="0" freq="10" grp="I-NP" id="261" key="1" loc="34.94" pos="DT" size="1" stem="a" suf="-">a</token>
    <token cog="0" freq="8" grp="O" id="262" key="295" loc="35.07" pos="``" size="2" stem="``" suf="->``</token>
    <token cog="0" freq="1" grp="I-NP" id="263" key="46" loc="35.21" pos="NN" size="7" stem="tracker" suf="-">tracker</token>
    <token cog="0" freq="4" grp="O" id="264" key="128" loc="35.34" pos="&#39&#39" size="2" stem="&#39&#39" suf="-">&#39&#39</token>
    <token cog="1" freq="2" grp="I-NP" id="265" key="22" loc="35.48" pos="NN" size="8" stem="comput" suf="er">computer</token>
    <token cog="0" freq="6" grp="I-NP" id="266" key="0" loc="35.61" pos="IN" size="4" stem="that" suf="-">that</token>
    <token cog="1" freq="1" grp="I-VG" id="267" key="33" loc="35.74" pos="VBZ" size="11" stem="coordin" suf="s">coordinates</token>
    <token cog="0" freq="50" grp="I-NP" id="268" key="4" loc="35.88" pos="DT" size="3" stem="the" suf="-">the</token>
    <token cog="0" freq="1" grp="I-NP" id="269" key="26" loc="36.01" pos="JJ" size="8" stem="match" suf="ing">matching</token>
    <token cog="0" freq="29" grp="O" id="270" key="6" loc="36.14" pos="IN" size="2" stem="of" suf="-">of</token>
    <token cog="1" freq="1" grp="I-NP" id="271" key="23" loc="36.28" pos="NNS" size="9" stem="consum" suf="er+s">consumers</token>
    <token cog="0" freq="4" grp="O" id="272" key="0" loc="36.41" pos="IN" size="4" stem="with" suf="-">with</token>
    <token cog="1" freq="1" grp="I-NP" id="273" key="24" loc="36.55" pos="NNS" size="9" stem="supplier" suf="er+s">suppliers</token>
    <token cog="0" freq="24" grp="O" id="274" key="2" loc="36.68" pos="." size="1" stem="." suf="->.</token>
</sentence>
```

Fonte: Produção do próprio autor.

As estruturas para as análises do corpo do texto são as *tags* personalizadas[3]: '<*paragraph*>', '<*sentence*>' e '<*token*>'. A *tag* '<*token*>' possui as seguintes propriedades:

a. *Cog* - indica se a palavra é cognata (valor '1') ou não (valor '0');
b. *Freq* - indica a frequência do item no texto;
c. *Grp* - indica a que grupo o item pertence: '*O*' (fora e classificação), '*I-NP*' (grupo nominal) e '*I-VG*' (grupo verbal);
d. *Id* - indica o índice da ordenação do item no texto;
e. *Key* - indica o valor de chavicidade do item no texto;
f. *Loc* - indica o valor em porcentagem da posição do item no texto (dispersão);
g. *Pos* - indica a etiqueta morfossintática do item;
h. *Size* - indica o tamanho do item em caracteres;
i. *Stem* - indica a forma do processo de *stemming* do item;
j. Suf - indica quais sufixos a palavra possui.

As análises em XML são lidas por um módulo que retorna informações requeridas por meio de funções. Uma das funções principais é a função '*filter_tokens(args,token_history=[])*', que retorna um conjunto de itens do texto a partir de uma série de condições passadas pelos seguintes argumentos:

a. *Limit* - limita o número de itens a ser retornado;
b. *Unique* - determina se os itens retornados podem ou não ser duplicados a partir dos valores '*yes*' ou '*no*';
c. *Minsize* - determina o tamanho mínimo do item em caracteres;
d. *Maxsize* - determina o tamanho máximo do item em caracteres;
e. *Stoplist* - uma lista de palavras que não podem ser incluídas no conjunto a ser retornado;
f. *Pos* - uma lista que determina quais etiquetas morfossintáticas de itens são aceitas para inclusão no conjunto a ser retornado;
g. *Stoppos* - uma lista que determina quais etiquetas morfossintáticas de itens não podem ser aceitas para a inclusão no conjunto a ser retornado;

[3] Criadas em XML pelo pesquisador para servir como uma metalinguagem.

h. *Punct* - determina se caracteres de pontuação podem ser retornados como itens a partir dos valores '*yes*' ou '*no*';
i. *Cognate* - determina a inclusão de palavras cognatas a partir dos valores '*yes*' ou '*no*';
j. *Minfreq* - determina a frequência mínima do item a ser incluído;
k. *Maxfreq* - determina a frequência máxima do item a ser incluído;
l. *Regexp* - determina um padrão de expressão regular para a aceitação de itens a serem retornados;
m. *Suffix* - determina o sufixo do item a ser validado para inclusão de itens a serem retornados;
n. *Min_position* - determina a posição mínima do item no texto para a aceitação de itens a serem retornados;
o. *Max_position* - determina a posição máxima do item no texto para a aceitação de itens a serem retornados;
p. *Repeat* - determina se itens já retornados em execuções anteriores da mesma função podem se repetir a partir dos valores '*yes*' ou '*no*';
q. *Word* - determina a inclusão de uma palavra específica;
r. *Order* - determina a ordem em que os itens são retornados: *keyness* (chavicidade), *keyness-asc* (chavicidade em ordem decrescente), *freq-asc* (frequência crescente), *freq-desc* (frequência decrescente), *word-asc* (ordem alfabética), *word-desc* (ordem alfabética decrescente), *position-asc* (posição crescente), *position-desc*(posição decrescente) e *random* ou *shuffle* (ordem aleatória/embaralhada).

A extração de informações de um *corpus*, previamente analisado pelas mesmas funções de análise de textos, é realizada por um módulo de análise de *corpora*. Uma das principais funções do módulo é a função que extrai sentenças a partir de uma palavra de busca ou padrão, conforme determinadas condições de argumentos. Os argumentos estão relacionados à quantidade de resultados e aos contextos da palavra de busca.

A criação de atividades a partir de informações extraídas de texto e *corpora*, são realizadas por *tags* personalizadas, tal como a *tag* '<*tokens*>'. As *tags* são combinadas em modelos que são interpretados para gerar atividades.

Os modelos em XML, criados a partir das análises automáticas programadas, que extraem informações de texto e *corpus* para a criação de atividades, permitem o reuso de funções que buscam dados linguísticos nos textos, uma vez que funcionam com uma

metalinguagem para a extração de tais dados linguísticos. Essa organização evita o trabalho de criação de códigos em linguagem de programação para cada modelo novo de atividade. A Figura 2 apresenta um exemplo de modelo de atividade em XML:

Figura 2 - Exemplo de modelo de atividade em XML

```
<activity>
    <label>Verifique o contexto gramatical das seguintes palavras no texto: </label>
    <label>[i]</label>
    <tokens limit='5'  pos='JJ' minsize='5'  case='lower' repeat='no' delimiter='; ' ></tokens>
    <label>[/i]</label>
    <line></line><line></line>
    <label>Responda conforme sua observação:</label>
    <line></line>
    <label>a. Qual a classe gramatical das palavras observadas?</label><line></line>
    <label>b. Geralmente, elas estão acompanhadas de que tipo de palavras?</label><line></line>
    <label>c. Quais outras palavras do mesmo tipo você consegue identificar no texto?</label>
    <line></line>
    <line></line>
</activity>
```

Fonte: Produção do próprio autor

Na figura, a linha de código '*<tokenslimit='5' pos='JJ' minsize='5' case='lower' repeat='no' delimiter='; ' ></tokens>*' pode ser traduzida na instrução a ser executada: a partir do texto de entrada analisado, retorne um limite de cinco itens que sejam adjetivos, com tamanho mínimo de cinco caracteres, em caixa baixa, sem repetições, delimitados pelo caractere ';'.

Um módulo específico interpreta os códigos em XML para a tradução dos modelos de atividade. Uma função faz a leitura de todas as *tags* no modelo e chama outras funções parametrizadas para a extração das informações requeridas. A decodificação do modelo na Figura 3 é apresentada a seguir:

Figura 3 - Exemplo de modelo interpretado

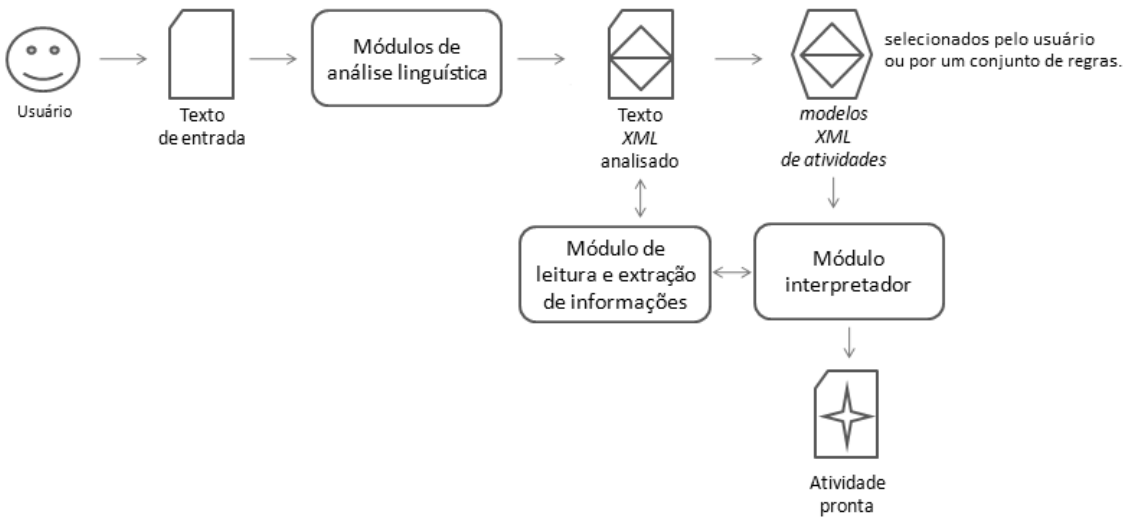
Verifique o contexto gramatical das seguintes palavras no texto: *mounting; vulnerable; negative; cocaine; daily;*

Responda conforme sua observação:
a. Qual a classe gramatical das palavras observadas?
b. Geralmente, elas estão acompanhadas de que tipo de palavras?
c. Quais outras palavras do mesmo tipo você consegue identificar no texto?

Fonte: Moreira Filho (2015)

A arquitetura completa do sistema é representada na Figura 4:

Figura 4 - Arquitetura do sistema.



Fonte: Moreira Filho (2015).

Na ilustração, são apresentados os passos do sistema para a criação de atividades. Primeiro, o usuário insere o texto de entrada, que será o foco para a criação das atividades de leitura. Em seguida, os módulos de análise linguística processam o texto e geram um conjunto de dados de análise armazenados em formato XML. Após as análises linguísticas, o programa realiza a seleção dos modelos de atividade a serem interpretados conforme as informações específicas do texto de entrada, tal como o gênero textual. Ao final, tendo interpretado os modelos, o sistema retorna à atividade pronta.

Os passos descritos fazem parte dos procedimentos de criação automática de atividades. Uma interface gráfica criada fornece opções adicionais que permitem ao usuário desativar ou ativar a sugestão de atividades automáticas e criar sua unidade didática a partir das análises e ferramentas disponíveis na interface.

A interface gráfica foi implementada por meio das linguagens *PHP*, *JavaScript*, *HTML* e *CSS*, além de componentes gratuitos de terceiros, como um editor de HTML, um conversor de códigos HTML para PDF e um criador de CAPTCHA[4]. A interface permite a visualização de análises automáticas do texto de entrada para uma análise pedagógica do potencial do texto para a criação de atividades. Um editor *HTML* possibilita a fácil edição das atividades elaboradas sem a necessidade de utilização de outros programas. Ao finalizar a

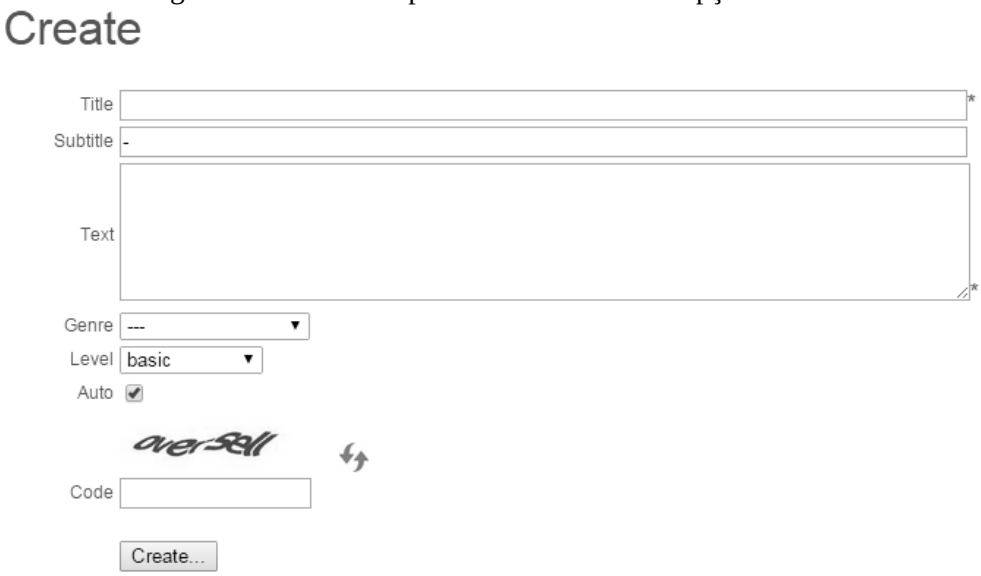
[4] Um tipo de teste cognitivo em páginas da Internet para diferenciar humanos de programas de computadores, a fim de evitar abusos e ataques a um sistema.

elaboração das atividades, é possível salvar a unidade didática em arquivo *PDF*, que fica armazenada no servidor da ferramenta para futuras buscas.

Um conjunto mínimo de páginas compõe a utilização básica do sistema para a criação de atividades: página inicial (*home*), página de busca de atividades criadas e salvas por usuários (*search*), página sobre o programa (*about*), página para a entrada do texto e opções do usuário (*create*), página de edição de atividades (*edit*) e página de visualização de atividades salvas (*saved*).

Para iniciar o processo de criação, é necessário acessar a página '*create*'. A Figura 5 apresenta o formulário a ser preenchido:

Figura 5 - Formulário para entrada de texto e opções do usuário.



Fonte: Moreira Filho (2015).

No formulário, é preciso preencher os campos obrigatórios com as informações do título do texto ('*Title*'), o corpo do texto ('*Text*'), além do campo opcional '*Subtitle*', e selecionar as opções também importantes para a preparação do material, como o gênero textual ('*Genre*'), o nível dos aprendizes ('*Level*') e a ativação do recurso para criação automática de uma unidade didática a partir do texto de entrada. É obrigatório também o preenchimento do código *CAPTCHA* para que o formulário seja enviado e as informações processadas.

Após o envio do formulário, o texto e as informações são processadas. A página de edição de atividades aparece com a guia editor ativada e seu conteúdo de acordo com as

opções definidas pelo usuário. A Figura 6 apresenta um exemplo de visualização da página de edição com a sugestão de atividades realizada pelo sistema:

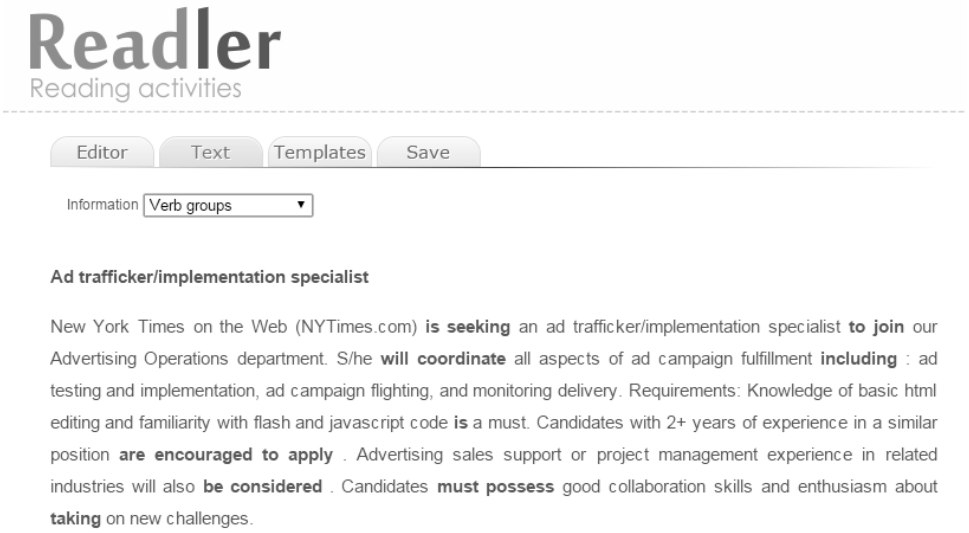
Figura 6 - Editor de atividades do Readler.

Fonte: Moreira Filho (2015).

A elaboração e a edição das atividades é realizada a partir de quatro abas: *Editor*, *Text*, *Templates* e *Save*, que possibilitam a execução das tarefas de edição, análise pedagógica do texto, seleção de modelos de atividades e publicação das atividades em unidades didáticas (o texto de entrada e as atividades elaboradas).

Na aba *Editor*, o usuário faz a edição das atividades criadas por meio de alterações simples de texto e formatação. A aba *Text* permite visualizar o texto de entrada e as análises linguísticas disponíveis, tal como os dados estatísticos sobre a frequência das palavras do texto, palavras cognatas, palavras-chave, grupos nominais e verbais, lista de frequência, afixos, lista de palavras-chave, frequência das classes morfossintáticas e informações de legibilidade do texto. A Figura 7 ilustra um dos tipos de visualização da aba *Text*:

Figura 7 - Exemplo de visualização de informações do texto na aba Text



Fonte: Moreira Filho (2015)

A figura exemplifica a identificação de grupos verbais no texto, visualizados em destaque com negrito. Na figura, também é possível perceber uma lista suspensa que é utilizada para selecionar o tipo de informação a ser exibida, como '*Verb groups*'.

A seleção de modelos disponíveis no sistema para criação e inclusão de atividades ao *Editor* é realizada na aba *Templates*. A Figura 8 a mostra um dos passos de utilização para criar uma atividade a partir da seleção de um modelo:

Figura 8 - Exemplo de seleção de modelos de atividades na aba *Templates.*

Fonte: Moreira Filho (2015).

A figura mostra a seleção de um dos modelos de atividade na caixa '*Items*'. Os modelos selecionados são inseridos na caixa '*Selection*'. Ao clicar em um dos modelos, uma pré-visualização do modelo de atividade interpretado em tempo real é exibida. Tal recurso facilita a decisão de inclusão dos modelos de atividade, uma vez que já é possível visualizar o resultado. O quadro a seguir ilustra alguns modelos interpretados, disponíveis na caixa '*Items*', a partir de análises linguísticas automáticas:

Quadro 2 - Questões do formulário de avaliação.

Análise linguística	**Modelo interpretado**
Identificação de palavras cognatas	Qual conjunto possui o maior número de palavras cognatas do texto? *( )evidence; substance; stress; neuroscience; depression; simulation; intimidating; molecular;* *( )yet; to; known; mood; that; such; are; as;* *( )more; the; also; eric; j.; makes; and; of;* *( )sinai; mount; at; new; school; his; nestler; he;*
Análise morfossintática	Busque as palavras abaixo no texto e as classifique em: adjetivo, substantivo ou verbo. a) disorders; b) negative; c) increase; d) abuse; e) given;
Identificação de afixos	Quais palavras são compostas por uma raíz mais afixo (prefixo ou sufixo): *subjected* *daily* *commonly* *artificially* *mounting* *vulnerable* *stressful* *shorter-than-usual*
Identificação de grupos verbais e nominais	Identifique os grupos com **V** (grupo verbal) ou **N** (grupo nominal): *( ) are known to increase* *( ) Mood disorders* *( ) to interact* *( ) intimidating mouse* *( ) to produce* *( ) this regulatory molecule* *( ) is believed to ratchet up* *( ) drug abuse*

Fonte: Moreira Filho (2015).

Além de modelos que são construídos a partir de análises linguísticas, são disponibilizados também enunciados e perguntas que não necessariamente dependem da extração de itens lexicais do texto de entrada. Como exemplo, podemos citar perguntas

relacionadas à identificação do gênero textual, perguntas gerais que consideram o texto como um todo e, também, enunciados para a síntese das informações do texto.

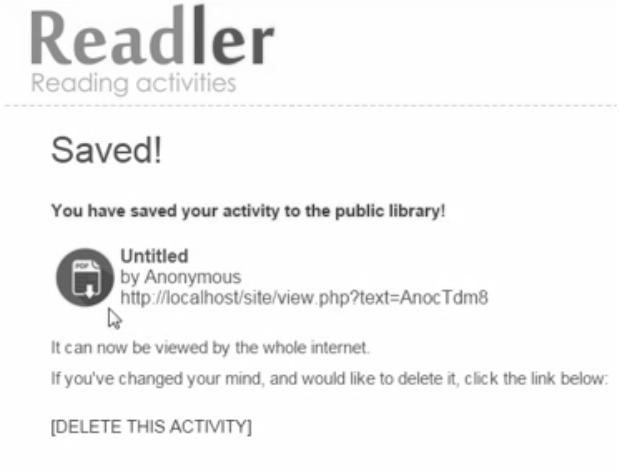
Para salvar as atividades elaboradas, utiliza-se a aba *Save*. A Figura 9 ilustra o formulário para publicação das atividades:

Figura 9 - Formulário para salvar atividades elaboradas.

Fonte: Moreira Filho(2015).

A publicação da unidade didática é realizada por meio do preenchimento e envio do formulário, com os campos '*Title*' (título da atividade), '*Author*' (identificação de quem está publicando) e '*Tags*' (palavras para facilitar a busca da atividade). A Figura 10 mostra um exemplo de mensagem de sucesso para a publicação em *PDF* da atividade:

Figura 10 - Mensagem da página após a publicação da atividade.



Fonte: Moreira Filho (2015).

Na mensagem, há também a opção de excluir a atividade publicada por meio de um *link* disponibilizado. A página que exibe a mensagem marca a etapa final de todo o processo de criação de atividades permitido pelo sistema desenvolvido.

6. Avaliação do sistema

A avaliação do sistema foi realizada em relação ao desempenho das análises linguísticas e aos possíveis usuários finais. Sobre as análises linguísticas, considerou-se que os módulos programados, de maneira geral, realizam satisfatoriamente as tarefas para as quais foram desenvolvidos, com resultados de acurácia entre 88% e 98% em tarefas específicas como identificação de palavras cognatas e etiquetagem morfossintática respectivamente.

O procedimento de avaliação, por usuários, contou com a participação de 20 professores de inglês da rede pública de ensino, os quais utilizaram a ferramenta e, em seguida, responderam um questionário com 16 perguntas, conforme o quadro a seguir:

Quadro 3 - Questões do formulário de avaliação.

Número	Questão
1	É difícil aprender a usar a ferramenta?
2	É fácil a utilização?
3	A ferramenta é eficiente no que deve fazer?
4	A ferramenta permite a elaboração de atividades de forma produtiva?
5	O processamento das tarefas é realizado de forma rápida?
6	Há erros no processamento das informações?
7	A ferramenta é flexível para realizar as tarefas de criação de atividades?
8	Qual seu grau de satisfação em relação às tarefas realizadas pela ferramenta?
9	Há itens ou partes da ferramenta que você não entendeu ou teve dificuldade de entender?
10	Os modelos e atividades disponíveis são pedagogicamente significativos para aprendizes da língua inglesa?
11	Como você avalia a criação automática das atividades pela ferramenta?
12	Qual o grau de utilidade da ferramenta para professores de língua inglesa?
13	As informações geradas pela ferramenta são confiáveis?
14	As análises são pertinentes e auxiliam no entendimento melhor do texto?

15	O que você sugere para melhorar a ferramenta?
16	Há alguma outra ferramenta que você utiliza para preparar atividades? Qual?

Fonte: Moreira Filho (2015).

As questões tiveram como objetivo avaliar as análises automáticas, o conjunto de modelos de atividades disponíveis e a interface do sistema. As questões fechadas foram disponibilizadas com uma escola de 1 a 5.

A avaliação do uso da ferramenta por possíveis usuários finais mostrou-se positiva tanto em relação a informações geradas pelo sistema, quanto a aspectos da interface.

7. Considerações Finais

Os resultados obtidos na avaliação foram considerados satisfatórios em relação aos objetivos e à motivação da pesquisa, embora haja a necessidade de avanços em relação ao sistema descrito, pois sugerem que ele apresenta um diferencial em relação a ferramentas disponíveis para análise de textos (concordanciadores), uma vez que fornece análises linguísticas variadas e interface desenhada para a tarefa de elaboração de atividades didáticas, quando comparado, por exemplo, a programas como concordanciadores, desenvolvidos para pesquisa em geral.

O estudo levanta evidências que suportam a viabilidade do uso de uma ferramenta computacional que aproveita os avanços científicos das áreas da Linguística e da Computação para a criação automática de atividades de leitura em língua inglesa e seu uso por professores, de forma a conferir funcionalidades úteis e confiáveis para a elaboração de atividades de ensino ao propiciar materiais que privilegiam a língua em uso. Além da criação de materiais de ensino, considera-se que a ferramenta pode também contribuir para a formação de docentes, tanto em relação à ampliação de seus conhecimentos linguísticos como ao seu aproveitamento na elaboração de atividades pedagógicas em sua prática.

Espera-se que o trabalho possa estimular novos estudos em relação à automatização das análises de exploração de *corpora* no desenvolvimento de ferramentas computacionais para o desenvolvimento de materiais de ensino e propiciar a ampliação das áreas de interesse da Linguística de Corpus no âmbito do ensino de línguas.

Recebido em: 02/2015; Aceito em 06/2015.

Referências Bibliográficas

BERBER SARDINHA, T. 2004. **Linguística de Corpus**. São Paulo: Manole.

BIRD, S.; LOPER, E.; KLEIN, E.2009.**Natural Language Processing with Python**. [S.l.]: O'Reilly Media Inc.

GRELLET, F. 1981.**Developing Reading Skills**. Cambridge: Cambridge University Press.

HOLMES, J. 1982. Stages, strategies, activities. **Working Papers**, n. 17, jul.

MOREIRA FILHO, J.L. 2015 **Linguística e Computação em diálogo para análise de textos e criação de atividades de leitura em língua inglesa**. 2015. 319 p. Tese (Doutorado) - FFLCH, Universidade de São Paulo, São Paulo, 2015.

RAMOS, R. C. G. 2004. Gêneros textuais: uma proposta de aplicação em cursos de inglês para fins específicos. **TheESPecialist**, v. 25, n. 2, p.107-129.

SCOTT, M.; CARIONI, L.; ZANATTA, M.; BAYER, E.; QUINTANILHA, T. 1984. Using a "standard exercise" in teaching reading comprehension. **ELT Journal**, v. 38, n. 2, p. 114-120.

José Lopes Moreira Filho is an English Teacher and Coordinator in the Education Secretary of São Paulo, Brazil, providing professional development for teachers. He has research experience in Linguistics, focusing on Applied Linguistics, acting on the following subjects: Corpus Linguistics, Language Teaching, Natural Language Processing and Machine Learning. E-mail: jlopes@usp.br