Oportunidades de uso da Adaptatividade em um SPLN para criação de atividades de leitura

J. L. Moreira Filho and Z. M. Zapparoli

***Resumo* — Este artigo apresenta oportunidades de utilização de adaptatividade em um sistema de processamento de língua natural (SPLN) para criação automática de atividades de leitura em textos com *corpora*. O trabalho consiste na construção de um sistema capaz de ler um texto em língua inglesa e automaticamente criar exercícios em forma de questões de múltipla escolha relacionadas ao texto de entrada, em um ambiente de virtual de aprendizagem próprio desenvolvido para tal fim. O principal foco do trabalho está relacionado à identificação, categorização e extração de itens lexicais para compor os exercícios das atividades. Primeiro, apresentamos brevemente alguns estudos prévios e protótipos de *software*. Em segundo, descrevemos as principais funcionalidades do ambiente virtual em que o SPLN. Em terceiro, discutimos as oportunidades de uso da adaptatividade para o sistema proposto. Por último, apresentamos as considerações finais.**

***Palavras-chave* — Linguística de Corpus (*Corpus Linguistics)*, Tecnologia Adaptativa *(Adaptative Technology)*, Processamento de Língua Natural *(Natural Language Processing)*.**

I. INTRODUÇÃO

Este trabalho apresenta alguns avanços obtidos no trabalho de pesquisa na criação de um SPLN para criação de atividades de leitura em língua inglesa com *corpora*.

Há uma gama de pesquisas sobre a análise, o desenvolvimento e o emprego de materiais de ensino baseados em *corpora*, uma vez que privilegiam a língua em uso[1].

Embora seja positivo para o processo de ensino-aprendizagem, o uso de materiais baseados em *corpora* ainda não é uma realidade comum fora do contexto acadêmico. Mesmo com as mais variadas ferramentas computacionais para análise de *corpora*, o processo de preparação de unidades didáticas inteiras e até mesmo de atividades pode ser considerado problemático para a maioria dos professores.

A tarefa, que geralmente leva tempo, é realizada apenas por pesquisadores; 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. Podemos citar, como exemplo, a pesquisa de [1], que descreveu todo o percurso do uso de dois *corpora* na elaboração de uma tarefa para ensino de inglês por meio de análises propiciadas por essas ferramentas.

Não é possível esperar que todo professor seja um especialista em Linguística de *Corpus* (LC) para que possa aproveitar os benefícios do uso de *corpus* em sala de aula.

Devido a esses motivos, professores podem ter dificuldades na preparação de tais materiais e, em consequência, não utilizá-los com certa frequência e/ou fazer uso de materiais tradicionais não significativos para a aprendizagem dos alunos.

Assim, a partir do desenvolvimento, aplicação e análise de um sistema de criação e montagem automática de atividades *online* de leitura em língua inglesa com *corpora*, por meio do uso de técnicas de análise de Processamento de Língua Natural (PLN), de práticas de análise de *corpus* para o ensino de línguas e do conceito de Adaptatividade e Tecnologia Adaptativa [2], a pesquisa em andamento tem o objetivo de suprir a necessidade de professores de língua estrangeira que desejam utilizar materiais baseados em *corpora*, mas que não estão familiarizados com o uso de ferramentas de processamento e exploração de *corpora* e/ou que não possuem muito tempo para preparar atividades.

A investigação está baseada em um estudo inicial realizado em uma pesquisa de mestrado [3], que teve como produto final um *software desktop* para preparação semiautomática de atividades de leitura em inglês.

Nos primeiros protótipos, com base no conceito de "atividade padrão" para o ensino de leitura de *English For Specific Purposes* (*ESP*), inglês para fins específicos, um conjunto fixo de exercícios é preparado automaticamente, incluindo atividades baseadas em concordâncias, *data-driven learning*[2], predição, léxico-gramática e questões para leitura crítica. Para tanto, o programa faz várias análises automáticas do texto selecionado por meio de fórmulas estatísticas: lista de frequência, palavras-chave, possíveis palavras cognatas, etiquetagem morfológica, possíveis padrões (*n-gramas*) e densidade lexical do texto.

Embora os resultados obtidos tenham demonstrado a viabilidade e o potencial de, por meio do computador, analisar textos e gerar automaticamente determinados tipos de exercícios para ensino de estratégias de leitura, há ainda a necessidade de muita pesquisa e desenvolvimento de melhorias para que a ferramenta possa ser usada pelo usuário final.

Assim, estudam-se propostas de utilização de adaptatividade no sistema em desenvolvimento.

II. ESTUDO INICIAL E PROTÓTIPOS

1. Primeiro Protótipo – Modelo Fixo

O *software desktop* desenvolvido na pesquisa de mestrado [3], para o sistema operacional Windows, na linguagem de programação Visual Basic 6, disponível no sítio http://www.xcorpus.net/downloads/rcb.zip, permite a criação de atividades a partir de modelos estáticos e de um assistente (*wizard*) que auxilia o usuário até a preparação das atividades.

[1] Textos reais, não inventados com o objetivo de ensinar língua.

[2] Proposta que enfatiza o ensino do léxico da língua por meio de descobertas, tornando o aluno um pesquisador.

O assistente eletrônico guia o usuário através de três etapas. Na primeira etapa, o usuário escolhe que tipo de atividade de leitura será preparado. O programa fornece alguns modelos padronizados. Na segunda etapa, o usuário seleciona um texto de sua escolha para preparação da atividade. Na terceira etapa, o programa exibe informações sobre o texto escolhido e a atividade quase pronta. Após verificar as informações do texto e editar a atividade (opcional), o usuário pode salvá-la.

Os exercícios das atividades estão relacionados a concordâncias (*data-driven*), predição, léxico-gramática e questões para leitura crítica. Um modelo fixo com questões e lacunas para inserção de itens lexicais do próprio texto são utilizadas para formar os exercícios na atividade.

Para tanto, a partir do texto selecionado, o programa realiza várias análises automáticas por meio de fórmulas estatísticas: lista de frequência, palavras-chave, possíveis palavras cognatas, etiquetagem morfológico, possíveis padrões (*n-gramas*) e densidade lexical do texto. O esquema a seguir sintetiza o funcionamento do sistema.

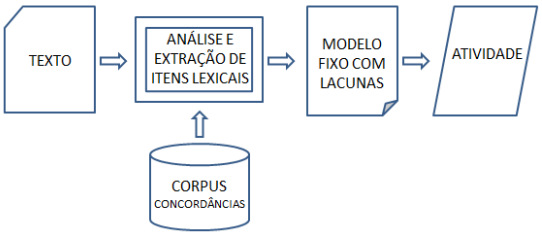

Figura 01 – Esquema de funcionamento do primeiro protótipo

2. Segundo Protótipo – Modelo Personalizável

Outra versão, tendo em vista os problemas de compatibilidade, foi escrita em C#, disponível no sítio: http://www.xcorpus.net/downloads/rcb_01-08-2010.zip, com o objetivo de manter o registro do trabalho realizado. O programa incorporou algumas mudanças em relação aos modelos utilizados para a criação das atividades e ao número de etapas.

O número de etapas do assistente foi reduzido a apenas três: Etapa 1 – escolha do modelo de atividade; Etapa 2 – escolha do texto e do *corpus* de referência, e extração de exemplos; Etapa 3 – visualização de informações do texto e atividade preparada para edição e publicação.

Os modelos podem ser criados pelo próprio usuário por meio da utilização de um conjunto finito de códigos, tal como <XC010>, que retorna dez possíveis palavras cognatas do texto, ou <XK0#>, que retorna um determinado número depor meio palavras-chave do texto (substituir # pelo número desejado). Na ajuda do programa, há uma listagem de todos os códigos disponíveis e sua descrição.

Nesse primeiro protótipo, com base no conceito de atividade padrão, introduzido em para cursos de ESP, embora os resultados obtidos tenham demonstrado a viabilidade e o potencial da solução, há ainda a necessidade de muita pesquisa e desenvolvimento para melhorias: diminuição de erros de análise, aumento da variedade de exercícios disponíveis e adequação do conjunto de exercícios ao texto de entrada.

Com base na experiência de desenvolvimento dessas duas versões, pretende-se explorar as possibilidades de montagem de atividades de forma automatizada, por meio de adaptatividade.

III. AMBIENTE VIRTUAL DE APRENDIZAGEM

O sistema de criação automática de atividade de leitura estará atrelado ao um ambiente virtual de aprendizagem, inicialmente chamado de '*Readler*', implementado em PHP (interface) e Python (processamento de língua natural). A figura abaixo ilustra a interface do ambiente *online*:

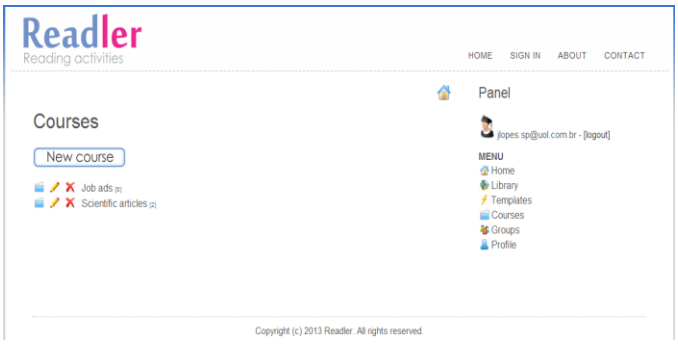

Figura 02 – Interface do ambiente virtual

O ambiente prevê a utilização de dois tipos de usuário: a. professor; b. aluno. Na interface, ao professor, são disponibilizadas ferramentas para a criação de cursos, atividades e grupos de alunos para os cursos criados. Ao aluno, são disponibilizados os cursos e suas respectivas atividades.

Para o professor, a utilização do ambiente consiste na manipulação dos seguintes principais itens do menu: *Library*, *Templates*, *Courses* e *Groups*.

A. *Library*

Seção em que é possível armazenar textos com o objetivo de servirem para a criação de atividades de leitura. É importante destacar que, antes de iniciar a criação de uma atividade, é necessário escolher um texto. O texto adicionado é analisado e armazenado em um banco de dados em formato XML para estar disponível para a criação de atividades em qualquer curso.

```
<sentence id="7">
        <token cog="y" freq="1" group="nI" id="155" key="7.0"
loc="77.9411764706" pos="NNS" pref="" suf="" syllables="4">Requirements</token>
        <token cog="n" freq="2" group="O" id="156" key="3.0"
loc="78.431372549" pos=":" pref="" suf="" syllables="0">:</token>
        <token cog="n" freq="6" group="O" id="157" key="79.0"
loc="78.9215686275" pos=":" pref="" suf="" syllables="0">-</token>
        <token cog="n" freq="1" group="nI" id="158" key="16.0"
loc="79.4117647059" pos="NNS" pref="" suf="" syllables="3">Bachelors</token>
        <token cog="n" freq="1" group="nI" id="159" key="6.0"
loc="79.9019607843" pos="NN" pref="" suf="" syllables="2">degree</token>
        <token cog="n" freq="1" group="nI" id="160" key="0.0"
loc="80.3921568627" pos="CC" pref="" suf="" syllables="1">or</token>
        <token cog="y" freq="1" group="nI" id="161" key="8.0"
loc="80.8823529412" pos="NN" pref="" suf="+ent" syllables="4">equivalent</token>
        <token cog="n" freq="11" group="O" id="162" key="1.0"
loc="81.3725490196" pos="SENT" pref="" suf="" syllables="0">.</token>
    </sentence>
```

Figura 03 – Exemplo de sentença de texto analisado no formato XML

As análises automáticas realizadas incluem: itemização, etiquetagem morfológica, contagem de frequência das palavras, extração de palavras-chave, identificação de palavras cognatas, identificação de grupos nominais, identificação de referência pronominal, identificação de afixos e marcação da posição de cada palavra no texto. Na mesma seção, há a possibilidade de o usuário visualizar as informações do texto em uma interface amigável.

B. Template

Seção em que o usuário tem a possibilidade de criar códigos em XML para a criação automática de atividades personalizadas, já que um conjunto padrão estará sempre disponível. Tal funcionalidade é considerada de nível avançado. Apresentamos abaixo um exemplo de *template* para criar uma questão de múltipla escolha com palavras cognatas.

```
<activity>
    <multiple_choice>
    <stem><label>What are the cognate words from the text?</label></stem>
    <opt_A>
    <text>a)</text>
    <cognates limit='5' delimiter=' ' sort ='text' order='asc' case='lower'></cognates>
    </opt_A>
    <opt_B><wordlist limit='5' delimiter=' ' sort ='text' order='asc' case='lower'></label></opt_B>
    <opt_C><wordlist limit='5:10' delimiter=' ' sort ='text' order='asc' case='lower'></opt_C>
    <opt_D><wordlist limit='10:15' delimiter=' ' sort ='text' order='asc' case='lower'></opt_D>
    <answer>A</answer>
    </multiple_choice>
</activity>
```

Figura 04 – Exemplo de código XML em um *template*

C. Courses

Seção para a criação de cursos. Os cursos são compostos por atividades. Cada atividade deve estar relacionada a um texto já adicionado à seção '*Library*'. Os itens que compõe as atividades são, a princípio, questões de múltipla escolha que podem ser criadas manualmente pelo usuário ou automaticamente por meio de '*templates*', já disponibilizados no ambiente ou personalizados.

D. Groups

Seção para o gerenciamento de alunos nos cursos criados. É possível organizar turmas e disponibilizar o conteúdo dos cursos, além de acompanhar o desempenho de cursistas.

Os passos básicos para a utilização do ambiente, na perspectiva do professor são: a. adição de textos à seção *Library*; b. criação de um curso na seção *Courses*; c. seleção de um curso para criação de uma atividade; d. criação de itens (manual ou automaticamente por meio de *templates*) para a atividade criada.

O ambiente *online* está em constante desenvolvimento, com adição de recursos de análises de texto e criação automática de atividades. Uma das intenções (posteriores) em relação ao sistema a ser implementado é a possibilidade de disponibilizar alguns recursos utilizados pelo professor para a interface do aluno, permitindo uma nova possibilidade de aprendizado conhecida como *self access language learning*, na qual o aluno teria ao seu dispor os recursos para escolha de textos e criação de suas próprias atividades.

Contudo, tanto em relação à utilização primária do ambiente como para uma possível estrapolação de seus objetivos, é necessário que as funções de criação automática de atividades sejam efetivas, de modo a retornar os melhores resultados possíveis sem erros ou inadequações.

IV. OPORTUNIDADES DE USO DE ADAPTATIVIDADE NO SPLN

O SPLN, desenvolvido em Python, para o ambiente virtual é composto basicamente de um módulo de análise linguística, a partir de técnicas e métodos de análises utilizados em processamento de língua natural e pesquisas em Linguística de *Corpus*, e um módulo com funções parametrizadas que faz a leitura e tradução de um texto analisado e códigos de *templates* em XML para criação de questões de múltipla escolha.

As principais análises linguísticas realizadas pelo módulo são: itemização, etiquetagem morfológica, geração de listas de frequência, extração de palavras-chave, identificação de palavras cognatas, identificação de grupos nominais e identificação de afixos.

Tendo em vista a natureza das análises e os objetivos explicitados em relação à criação de atividades, é imprescindível que erros e imprecisões sejam eliminados.

Muitas das análises e heurísticas utilizadas no módulo, principalmente as derivadas de metodologias da instrumentação de análise de *corpus*, são baseadas unicamente em estatísticas. Por exemplo, para a extração de palavras-chave, há a comparação das frequências das palavras no texto com uma lista de frequência de um banco de dados de quase 100.000.000 de palavras por meio de uma fórmula (*log likelihood*), com o retorno de uma lista de palavras ordenadas por chavicidade. Os resultados geralmente precisam ser filtrados a fim de se obter uma listagem sem ruídos.

O exemplo ilustra um problema comum com processos estocásticos, embora sua utilização seja ampla no processamento de língua natural. A obtenção de dados exatos por sua natureza não determinística torna-se insuficiente. Da mesma maneira, métodos unicamente gramaticais e simbólicos podem não bastar a determinados tipos de problema.

É o caso do método utilizado na identificação de grupos nominais por meio de uma gramática, no módulo de análise do SPLN. O método de [4] utiliza três etiquetas básicas (*inside (nI)*, *outside (O)* e *between(nB)*) e leva em consideração a etiqueta morfológica dos itens. Uma primeira etiquetagem é feita e uma gramática faz o refinamento da etiquetagem por meio de regras contextuais. Um autômato simples faz a leitura e identificação dos grupos nominais, como mostra a figura abaixo:

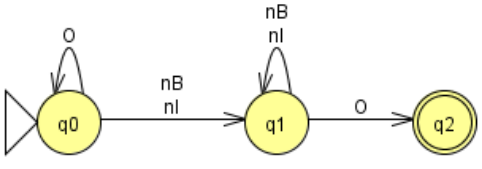

Figura 05 - Exemplo de autômato para leitura de grupos nominais

Para a obtenção de melhores resultados, a combinação de métodos gramaticais e autômatos com processos estocásticos é desejável. O uso de soluções híbridas pode ser a chave para resolução de problemas quando uma única tecnologia empregada não dá conta de toda a complexidade. No caso do processamento de língua natural, em que o problema de ambiguidades é recorrente, é justificável a busca de conciliação de métodos aparentemente antagônicos.

O elo para a conciliação desses métodos pode ser o uso da adaptatividade. A aplicação da adaptatividade pode permitir pontos de equilíbrio entre métodos simbólicos e estocásticos. Podemos citar o trabalho de [5], que estuda um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos.

No SPLN em desenvolvimento, um problema interessante que engloba um conjunto amplo de análises linguísticas já mencionadas e que pode ampliar as funções de criação de atividades é a localização e extração de informações do texto. Em determinados gêneros textuais, como, por exemplo, anúncios de emprego ou artigos científicos, os quais ambos formam os *corpora* utilizados no desenvolvimento do sistema, há padrões de informações que podem ser identificados a fim de evidenciar a estrutura interna do gênero/texto e, por conseguinte, como fonte para a formação de questões.

Nos textos do *corpus* de anúncios de emprego, as seguintes informações podem ser identificadas: quem/qual empresa oferece a vaga, que tipo de vaga é oferecido, quais são as exigências e qualificações necessárias para a vaga, qual o salário e os benefícios oferecidos e como proceder para se candidatar à vaga. Padrões recorrentes como '*We are seeking a…*', '*We offer a competitive salary…*' e '*Send resume to…*' podem funcionar como índices para extração dessas informações.

A princípio, uma função com expressões regulares ou uma gramática poderia ser utilizada para a extração das informações. Porém, no caso, há uma grande possibilidade de implementação de adaptatividade, tendo em vista a natureza do problema, as diferentes variáveis envolvidas, podendo ser aplicadas gramáticas e autômatos em conjunto com dados estatísticos.

Para tanto, faz-se a necessidade de uma avaliação e análise do que pode ser conseguido por meio de métodos simbólicos e estocásticos para a implementação de uma solução híbrida, com possibilidades de características de aprendizagem de máquina em caso de extensão das funcionalidades para textos de diferentes gêneros.

V. Considerações finais

O trabalho buscou mostrar alguns dos avanços feitos em uma pesquisa no desenvolvimento de um SPLN para a criação de atividades de leitura em língua inglesa com corpora, com vistas à utilização da adaptatividade na melhoria dos resultados obtidos em análises linguísticas para a ampliação das funcionalidades do sistema proposto.

As informações apresentadas serão desdobradas na exploração do uso da adaptatividade em cada oportunidade de sua aplicação na conciliação dos diferentes métodos discutidos.

Ao final, espera-se a implementação da adaptatividade no sistema, a fim de consolidar o uso de tal tecnologia no projeto proposto, visando a uma contribuição interdisciplinar nas áreas da Linguística e Tecnologia Adaptativa.

Referências

[1] CONDI, R. Dois corpora, uma tarefa. O percurso de coleta, análise e utilização de corpora eletrônicos na elaboração de uma tarefa para ensino de inglês como Língua Estrangeira. 2005. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem), LAEL, PUC-SP, São Paulo, 2005.

[2] DIZERÓ, W. Formalismos Adaptativos Aplicados na Modelagem de Softwares Educacionais. 2010. Tese (Doutorado em Engenharia Elétrica e Sistemas Digitais), EPUSP, São Paulo, 2010.

[3] MOREIRA FILHO, P. Desenvolvimento de um software para preparação semiautomática de atividades de leitura em inglês. 2007. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem), LAEL, PUC-SP, São Paulo, 2007.

[4] RAMSHAW, L. AND MARCUS, M. P. (1995). Text chunking using transformation-based learning. In Yarowsky, D. and Church, K., editors, Proceedings of the Third Workshop on Very Large Corpora, pages 82–94, Cambridge, Massachusetts.

[5] MENEZES, C. E. D. ; JOSÉ NETO, J. . Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos.. In: Conferencia Iberoamericana en Sistemas, Cibernética e Informática - CISCI 2002, 2002, Orlando. Anais da Conferencia Iberoamericana en Sistemas, Cibernética e Informática - CISCI 2002., 2002.

José Lopes Moreira Filho é Doutorando em Semiótica e Linguística Geral (USP). Possui Mestrado em Linguística Aplicada e Estudos da Linguagem pela Pontifícia Universidade Católica de São Paulo (PUCSP). Possui graduação em Letras – Português e Inglês (Bacharelado Tradução) pela Universidade de Mogi das Cruzes (UMC). Atualmente, é Professor Coordenador do Núcleo Pedagógico da Diretoria Regional de Ensino Leste 3 da SEE-SP, mantendo interesses na área de Linguística, Linguística Aplicada, Linguística Informática, Linguística de *Corpus*, Processamento de Linguagem Natural, atuando principalmente no desenvolvimento de ferramentas computacionais para exploração de *corpora*, ensino de línguas, entre outras aplicações que envolvem linguagem e tecnologia.

Zilda Maria Zapparoli é professora associada aposentada junto ao Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, instituição em que obteve os títulos de Mestre, Doutor e Livre-Docente, e onde continua desenvolvendo atividades de ensino, pesquisa e orientação no Curso de Pós-Graduação em Linguística, área de Semiótica e Linguística Geral, linha de pesquisa Informática no Tratamento de *Corpora* e na Prática da Tradução. Desde 1972, atua em Linguística Informática, com tese de doutorado, tese de livre-docência, pós-doutorado na Université de Toulouse II e trabalhos publicados na área. É líder do Grupo Interdisciplinar de Pesquisas em Linguística Informática, certificado pela USP e cadastrado no Diretório de Grupos de Pesquisa no Brasil do CNPq, em 2002. Integrou comissões e colegiados na USP, destacando-se os trabalhos relativos ao processo de informatização da FFLCH-USP, enquanto membro da Comissão Central de Informática da USP e presidente da Comissão de Informática da FFLCH-USP por cerca de treze anos.