Adaptatividade para pesquisas em Linguística de *Corpus*

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

***Abstract* — This paper introduces some of the main computational tools, methodology and problems for language teaching research in *Corpus* Linguists, attempting to argue for the need of investigation in the use of Adaptive Technology in the area as an improvement. In order to illustrate the viability and relevance of the proposal, examples are also presented and discussed. It may be considered as a starting point for future works on both areas.**

***Keywords* — *Corpus* Linguistics, Adaptative Technology, computational tools.**

I. INTRODUÇÃO

Este trabalho apresenta algumas das principais ferramentas computacionais, metodologia e problemas que pesquisadores na área de Linguística de *Corpus*, especificamente no campo de atuação do ensino-aprendizagem de língua estrangeira, enfrentam em suas pesquisas, na tentativa de argumentar sobre a necessidade de investigação do uso do conceito e técnicas adaptativas na área a fim de buscar soluções que possam avançar as práticas correntes.

Primeiro, destacamos a importância do uso de *corpus*, textos autênticos, no ensino de línguas, motivo pelo qual há grande interesse de utilização do material de *corpus* em pesquisas para elaboração de cursos, material pedagógico, avaliação de livros didáticos existentes, desenvolvimento de *software*, entre outras aplicações.

Em seguida, apresentamos brevemente as áreas da Linguística de *Corpus* e Tecnologia Adaptativa. Uma descrição um pouco mais detalhada sobre a Linguística de *Corpus* é fornecida, uma vez que será o foco de análise para as discussões que seguirão.

Logo após, para ilustrar as ferramentas e práticas correntes no contexto mencionado, são apresentadas e descritos os principais programas e a metodologia básica para pesquisas em Linguística de *Corpus*, principalmente no âmbito do ensino-aprendizagem de línguas,

Tendo fornecido uma visão geral das características do instrumental para as pesquisas, argumentamos sobre a necessidade de investigação da possibilidade de uso do conceito e técnicas adaptativas a partir de problemas e lacunas na prática de pesquisa em Linguística de *Corpus*.

Como fechamento, apresentamos as considerações finais e possíveis caminhos para futuros trabalhos em ambas as áreas.

II. A IMPORTÂNCIA DO USO DE TEXTOS AUTÊNTICOS

A utilização de textos autênticos, textos que não foram inventados para ensinar língua, é vital se o objetivo de ensino é a comunicação, a interação social e a execução de tarefas no mundo real [1]. A utilização de textos autênticos em sala de aula é benéfica para aprendizagem.

Entre os argumentos a favor dessa utilização está o de que ela promove um aumento da motivação, pois o aluno sente que está aprendendo a "língua de verdade".

Conforme [2], muitos livros didáticos, por uma série de questões, utilizam textos inventados para o ensino de línguas, o que pode trazer efeitos negativos para a aprendizagem: desmotivação, habilidades não plenamente desenvolvidas, e uso não natural da língua.

O trabalho com textos autênticos e o uso de *corpus* para fins pedagógicos na escola podem trazer alternativas significativas para a aprendizagem de língua. Com eles, é possível dar condições para que os alunos se conscientizem sobre as unidades pré-fabricadas da língua, que são os padrões léxico-gramaticais, e a característica probabilística da linguagem, isto é, como a língua realmente funciona.

Os exemplos extraídos de *corpus* são importantes para a aprendizagem de línguas porque expõem os alunos, desde os estágios iniciais do processo de aprendizagem aos tipos de frases e vocabulários que possivelmente serão encontrados em textos autênticos da língua ou no uso da língua em situações reais de comunicação.

Para isso, existe uma ampla instrumentação que pode ser aproveitada por professores e pesquisadores na área de ensino de línguas. Há uma série de *corpora* e ferramentas computacionais sendo desenvolvidas, ainda não totalmente conhecidas e utilizadas por esses profissionais, especificamente professores.

Contudo, é preciso verificar se as ferramentas disponíveis são acessíveis e adequadas para profissionais em diferentes contextos, refletindo sobre a viabilidade de sua introdução. A disponibilidade de toda a instrumentação é desejada, o que pode levar a questões de adaptação e criação de soluções de fácil uso.

III. TECNOLOGIA ADAPTATIVA

Conforme [3], a Tecnologia Adaptativa está relacionada a técnicas, métodos e disciplinas que estudam as aplicações da adaptatividade, que pode ser entendida como uma propriedade que um determinado modelo tem de modificar espontaneamente seu próprio comportamento em resposta direta a uma entrada, sem auxilio externo.

Um sistema adaptativo é aquele que possui a propriedade de se auto modificar a partir de determinada entrada, sem a necessidade de um agente externo.

Dentro da Tecnologia Adaptativa, há a noção de dispositivo, uma abstração formal. O dispositivo pode ser adaptativo ou não adaptativo. O dispositivo não adaptativo pode ser formado por um conjunto finito de regras estáticas

que, em linguagem de programação, pode ser representado na forma de cláusulas IF-THEN. A operação do dispositivo se dá pela aplicação das regras, tendo como retorno determinados estados. Quando o dispositivo não aplica nenhuma regra, a operação é terminada, gerando um erro. As ações de dispositivos adaptativos podem ser chamadas quando ocorre algum erro (quando nenhuma regra é aplicável), ou quando a operação do dispositivo não adaptativo está em um determinado estado.

Basicamente, os dispositivos adaptativos são formados por três ações adaptativas elementares [4]: i. Consulta de regras/estados; ii. Exclusão de regras; iii. Inclusão de regras. Seu uso está ligado a situações complexas em que há a necessidade de tomadas de decisões não triviais, por exemplo, na área de estudos da linguagem, resolução de ambiguidades em programas de anotação (morfológica, sintática, etc.).

IV. Linguística de *Corpus*

A Linguística de *Corpus* é uma área que estuda a língua por meio da observação de grandes quantidades de dados linguísticos reais, isto é, textos falados ou escritos provenientes da comunicação no mundo real língua em uso), com o auxílio de ferramentas computacionais.

De forma geral, o conjunto de dados linguísticos reais criteriosamente coletados e utilizados em estudos de Linguística de *Corpus* é chamado de *corpus* (plural: *corpora*). O *corpus* deve ser constituído de dados autênticos (não inventados), legíveis por computador e representativos de uma língua ou variedade da língua que se deseja estudar.

A Linguística de *Corpus* faz uso de uma abordagem empirista e tem como central a noção de linguagem enquanto sistema probabilístico. De acordo com essa noção, os traços linguísticos não ocorrem de forma aleatória, sendo possível evidenciar e quantificar regularidades (padrões). É comum na área afirmar que a linguagem é padronizada (*patterned*), isto é, existe uma correlação entre os traços linguísticos e os contextos situacionais de uso da linguagem.

Na Linguística de *Corpus*, a padronização evidencia-se por colocações, coligações ou estruturas que se repetem significativamente

Algumas das áreas de interesse da Linguística de *Corpus* são: compilação de *corpora* e desenvolvimento de ferramentas para sua análise, descrição de linguagem, exploração do uso de descrições baseadas em *corpora* para várias aplicações, tal como o ensino de línguas, processamento de linguagem natural, reconhecimento de voz e tradução.

O computador desempenha um papel importante para os estudos na área, que foi impulsionada por seus avanços. As ferramentas computacionais são geralmente utilizadas para reorganização e extração de informações no *corpus* para observação e interpretação de dados, fornecendo novas perspectivas para a análise linguística.

Por meio das ferramentas computacionais, as pesquisas linguísticas podem ganhar velocidade, precisão e confiabilidade. Há uma comparação muito comum na literatura segundo a qual o computador e seus recursos seriam para o desenvolvimento da Linguística equivalentes ao do microscópio para a biologia.

Assim, surge uma grande variedade de *software* disponível para os estudos de Linguística de *Corpus*. Muitos *softwares* exibem e permitem a manipulação de extensas listas de frequência de palavras, fraseologias e colocações.

[5] apresenta uma série de ferramentas computacionais que podem ser utilizadas em pesquisas de Linguística de *Corpus*, as quais serão citadas, brevemente, em três categorias: i. programas para coleta de *corpus*; ii. Etiquetadores; iii. programas para análise da padronização.

V. Principais ferramentas utilizadas em pesquisas de Linguística de *Corpus*

Programas para coleta de *corpus*, como o *WinHTTrack*, um *off-line browser*, disponível em http://www.httrack.com/, permite o download de um site inteiro da Internet para o diretório de um computador local, incluindo HTML, imagens e outros tipos de arquivos. Em suas opções avançadas, é possível especificar tipos de arquivos e condições de busca para a coleta. Embora o programa não seja destinado especificamente para a pesquisa linguística, suas funções vão ao encontro da necessidade da coleta de grandes quantidades de textos na área, a qual também pode ser conseguida por meio de outros programas e *scripts* de linguagem de programação.

Outro tipo de programa, geralmente necessário após uma coleta automática de textos na Internet, é um conversor de HTML para textos sem formatação e códigos de script característicos. Tais programas podem ser encontrados facilmente na Web.

Também é possível escrever *scripts* em linguagem de programação Shell ou Python, por exemplo, que utilizem expressões regulares para a remoção de determinados códigos e textos indesejados, fazendo uma 'limpeza' nos textos do *corpus* para o prosseguimento das análises. Para os usuários de Windows, uma opção é a instalação do Cygwin, um emulador do sistema operacional Unix para Windows. Os scripts podem ser escritos e executados em um terminal de comandos.

Etiquetadores online e desktop, tais como o etiquetador do site VISL, Brill e TreeTagger, são utilizados para inserir automaticamente etiquetas que podem indicar, dependendo do tipo de etiquetagem, marcações morfossintáticas, sintáticas, semânticas ou discursivas. Um exemplo de tais marcações é ilustrado a seguir:

```
( [(] PU @PU #1->0
Reuters [Reuter] N P NOM @APP #2->0
) [)] PU @PU #3->0
- [-] PU @PU #4->0
The [the] ART S/P @>N #5->7
Arctic [Arctic] ADJ POS @>N #6->7
zone [zone] N S NOM @SUBJ> #7->8
has [have] <aux> V PR 3S @FS-ACC> #8->38
moved [move] <mv> V PCP2 AKT @ICL-AUX< #9->8
into [into] PRP @<SA #10->9
a [a] ART S @>N #11->12
warmer [warm] ADJ COM @P< #12->10
, [,] PU @PU #13->0
greener [green] ADJ COM @>N #14->19
« [«] PU @PU #15->0
new [new] ADJ POS @>N #16->19
normal [normal] ADJ POS @>N #17->19
» [»] PU @PU #18->17
phase [phase] N S NOM @SUBJ> #19->0
, [,] PU @PU #20->0
which [which] <rel> INDP S @SUBJ> #21->22
means [mean] <mv> V PR 3S @FS-N< #22->19
less [little] DET COM S @>N #23->24
```

Fig. 1. Exemplo de texto etiquetado

A maioria das ferramentas disponíveis permite uma utilização avançada, em que o conjunto de etiquetas pode ser aumentado e/ou o usuário pode fazer um treinamento dos dados para melhor satisfazer as necessidades específicas de sua pesquisa.

Programas para análise da padronização, como listadores de palavras e concordanciadores, que são programas que permitem a busca por palavras específicas em um *corpus*, fornecendo exaustivas listas para as ocorrências da palavra em contexto.

Geralmente esses programas são encontrados em uma única solução/*software*. É o caso de *software* como o famoso Wordsmith tools, com versões disponíveis para download em http://www.lexically.net/wordsmith/, um dos mais utilizados em pesquisas de Linguística de *Corpus*, cuja versão mais estável possui três ferramentas principais: Wordlist (lista palavras por frequência no *corpus*), KeyWords (extrai palavras-chave do *corpus*) e Concord (gera linhas de concordância).

Outros *software* que executam funções similares são: MicroConcord (http://www.lexically.net/software/index.htm), AntConc (http://www.antlab.sci.waseda.ac.jp/software.html), Unitex (http://www-igm.univ-mlv.fr/~unitex/) e Kitconc (http://www.fflch.usp.br/dl/li/x/?p=435), uma versão desktop simples em português.

Vejamos a interface de um programa que executa a função de contagem da frequência de palavras em corpora:

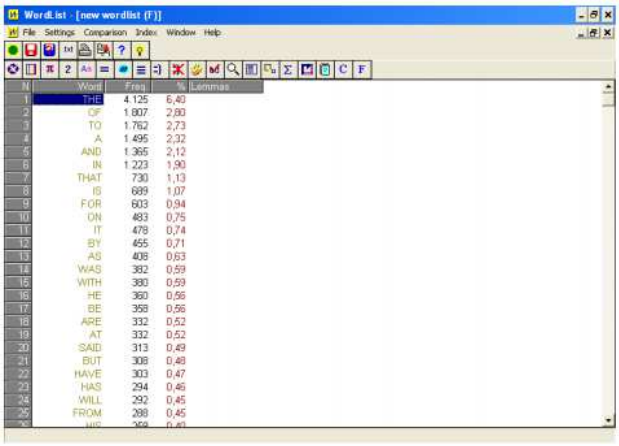

Fig. 2. Lista de frequência de palavras no Wordsmith tools 3.0

A partir de listadores de palavras, é possível descobrir quais são as palavras mais utilizadas em um texto ou conjunto de textos. Uma das opções permite listar também as expressões mais ocorrentes formadas por duas ou mais palavras, chamadas de n-gramas de um texto ou conjunto de textos.

A lista de palavras é uma listagem ordenada por frequência de todas as formas que ocorrem em um *corpus*. A partir da lista de frequência podemos definir quais são as palavras mais interessantes para análise do *corpus*.

A ideia é a de que palavras que possuem uma ocorrência maior são mais relevantes, visto que há uma probabilidade maior de serem encontradas em outro *corpus* ou textos. Por exemplo, para um aprendiz inicial de língua estrangeira, aprender palavras mais frequentes é extremamente essencial, visto que podem aparecer em diversos contextos.

Ao analisar a ocorrência das palavras gramaticais, podemos tentar identificar quais palavras se destacam em relação ao tipo de *corpus*, texto, gênero ou registro as quais pertencem. Geralmente, a palavra mais frequente (número um da lista) em textos em língua portuguesa é a preposição 'de'. Se alguma outra palavra gramatical ocupar esta posição, será uma ocorrência marcada e merecedora de verificação.

A manipulação das listas por meio de funções de classificação permite, quando o pesquisador julgar necessário, juntar frequências de palavras, como é o caso da lematização. No exemplo abaixo, o pesquisar poderia ter a necessidade de juntar todas as frequências das diferentes formas do verbo 'ABRIR' sob uma única forma (infinitivo):

3	ABERTO	2	00,02		
4	ABRA	4	00,05		
5	ABRE	3	00,04		
6	ABRIR	4	00,05		

Fig. 3. Exemplo de palavras marcadas para lematização

Podemos também levantar informações estatísticas gerais do *corpus*, como número total de palavras/itens (*tokens*), número de vocábulos/formas (*types*), razão forma/item (*TypeToken Ratio*), que pode ser calculada da seguinte maneira: ((número total de formas / número de itens)*100). A porcentagem resultante pode indicar a proporção de repetição das palavras no *corpus*. Um valor baixo da razão forma/item indica que há muita repetição no *corpus*, ou seja, um número menor de formas. Um valor alto da razão forma/item indica pouca repetição das palavras no *corpus*, um número mais de formas.

Outra maneira de fazer um recorte em relação às palavras que devem ser analisadas é a extração de palavras-chave. Muitas vezes, a lista de palavras-chave fornece uma filtragem mais apurada das palavras que se destacam em *corpus* ou texto.

Em um *corpus* formado por textos de anúncios de emprego em inglês, palavras como '*EXPERIENCE*', '*SKILLS*', '*ABILITY*', '*REQUIREMENTS*' e '*SALARY*' seriam palavras-chave, uma vez que podem ser consideradas especificas do gênero, revelando parte de sua estrutura padronizada:

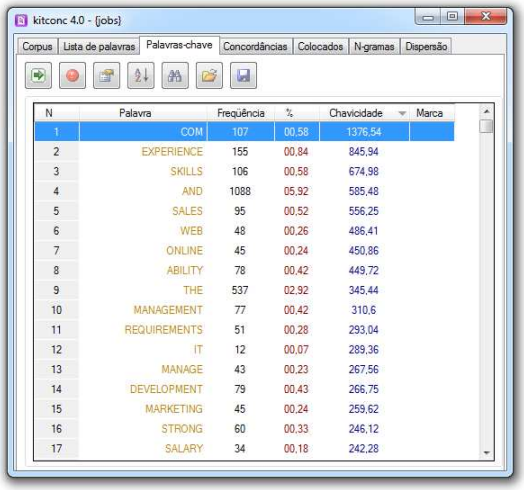

Fig. 4. Palavras-chave de no Kitconc 4.0

Em programas que executam extração de palavras-chave, é comum haver um recurso que possibilita a comparação automática de uma lista de frequência de palavras do *corpus* de estudo com uma lista de frequência de palavras de um *corpus* de referência, geralmente muito maior, a partir de um *corpus* de língua geral.

Em programação, tal comparação pode ser feita utilizando uma fórmula como mostra o *script* abaixo:

```
def loglikelihood(self,ntokensinsc,ifreqiinsc, ntokensincr,ifreqincr):
    #convert to float
    nt1 = float(ntokensinsc)#number of tokens in study corpus
    nt2 = float(ntokensincr)#number of tokens in ref corpus
    f1 = float(ifreqiinsc)#freq of item in study corpus
    f2 = float(ifreqincr) #freq of item in ref corpus
    #calculate
    e1 = nt1 * ((f1+f2) / (nt1+nt2))
    e2 = nt2 * ((f1+f2) / (nt1+nt2))
    ll = 2 *  ( (f1 * math.log(f1/e1) ) + (f2 * math.log(f2/e2) ) )
    #return value
    return ll
```

Fig. 5. Função *loglikelihood* para extração de palavras-chave

A função é aplicada para cada palavra na lista de frequência. Os argumentos da função, em ordem, são: frequência da palavra no *corpus* de estudo, número total de *tokens* no *corpus* de estudo, número total de *tokens* no *corpus* de referência, frequência da palavra no *corpus* de referência.

O resultado da comparação retorna uma lista classificada em que as principais palavras do *corpus* de estudos estarão no topo, geralmente palavras de conteúdo, em contraponto a uma lista de frequência, em que as primeiras palavras são gramaticais.

Uma concordância é a 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).

Em um primeiro momento, é comum associar/confundir o termo concordância apresentado aqui com o termo utilizado em gramática.

Quando utilizamos o termo concordância, queremos nos referir a um tipo de visualização privilegiada do uso de palavras, conforme já descrita. No exemplo abaixo, exibimos linhas de concordância da palavra 'tendência' em seu formato mais comum, KWIC (*key word in context*).

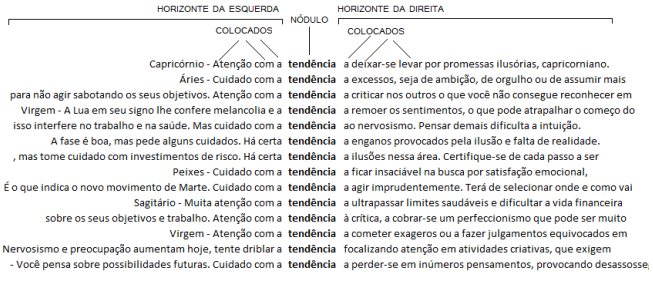

Fig. 6. Exemplo de linhas de concordância

Diferentemente de um texto normal, as linhas de concordâncias são geralmente lidas a partir de seu nódulo, palavra ou expressão de busca, a qual fica centralizada para análise. Na leitura, buscamos a existência de padrões de uso por meio da análise de colocados, palavras à esquerda e à direita que ocorrem significativamente com o nódulo. Por isso, em primeiro, não aconselhamos ficar buscando a totalidade dos fragmentos, a fim de completar a frases/sentenças.

Como exemplo, podemos primeiramente buscar quais palavras da esquerda e da direita geralmente se colocam com a palavra de busca a fim de determinar padrões. Em seguida, podemos analisar quais tipos de palavras formam tais padrões, classe gramatical, campo semântico, etc. Por fim, são identificados os sentidos de cada padrão. Vejamos a padronização da palavra '*resume*' nas linhas de concordância a seguir:

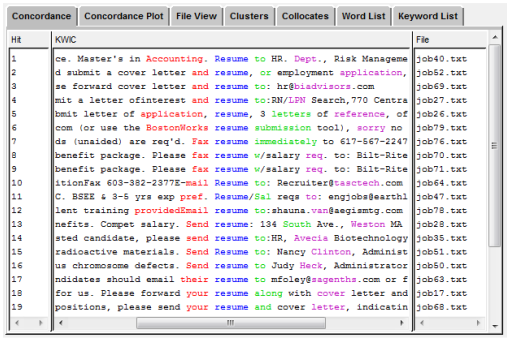

Fig. 7. Concordâncias da palavra '*resume*' no AntConc

Podemos analisar linhas de concordância e identificar três tipos de padrões: colocação, coligação e prosódia semântica. Colocação refere-se à junção significativa, co-ocorrência, entre palavras (*send*/*fax* + *resume* + *to*). Coligação é a relação gramatical mantida pelas palavras (verbo + resume + preposição). Prosódia semântica negativa, positiva ou neutra (o padrão *send/fax resume to* ocorre tipicamente com palavras de semântica e contextos neutros).

A identificação de padrões pode ser auxiliada também por outros tipos de visualização, como lista de colocados:

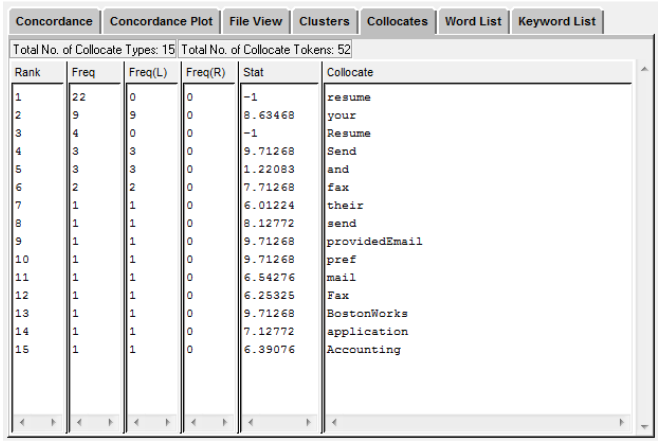

Fig. 8. Lista de colocados da palavra *resume* no AntConc

Na lista de colocados, podemos identificar a ocorrência das palavras '*send*' e '*fax*', colocados que formam o padrão identificado.

Quando há a necessidade de uso de algum recurso não encontrado nessas ferramentas para manipulação de seus dados, é comum a utilização de programas complementares.

Um desses programas pode ser a planilha eletrônica do Excel, que disponibiliza opções de filtro, classificação, cálculos de fórmulas, entre outras. Vejamos a filtragem de substantivos, identificados pela etiqueta 'NN', a partir de um texto etiquetado:

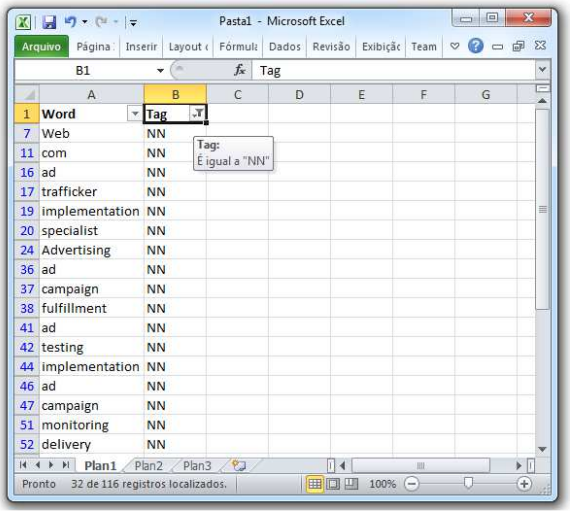

Fig. 9. Utilização do recurso de filtro em dados anotados

Embora o exemplo de utilização pareça simples, em conjunto com outras funções, o aplicativo pode suprir necessidades de determinadas pesquisas.

Um exemplo é a pesquisa de [6], que comparou os resultados de diferentes ferramentas de análise de *corpus* na tarefa de identificação de palavras candidatas a termo em textos sobre câncer de mama e desenvolveu uma metodologia com o auxílio do Excel para a filtragem dos candidatos mais comuns entre todas as ferramentas e candidatos exclusivos de cada uma.

VI. Uso das ferramentas computacionais em metodologia básica para pesquisas

As ferramentas descritas anteriormente podem ser aplicadas nos seguintes passos em pesquisas, compondo uma metodologia básica, em nível menos profundo, para a elaboração de cursos e extração de material para confecção de atividades pedagógicas. É possível que os passos descritos também sejam utilizados em pesquisas com diferentes objetivos além dos especificados, tal como lexicografia análise de *corpora* de aprendizes.

Vejamos os seguintes passos:

- Coleta do *corpus*;
  - Manual ou automática a partir de programas específicos.
- Preparação do *corpus* para as análises;
  - Limpeza e organização dos textos.
- Análise das frequências das palavras do *corpus*;
  - Busca de indícios de palavras que se destacam no *corpus*.
- Análise de palavras-chave do *corpus*;
  - Busca de palavras que se destacam no *corpus*.
- Seleção de palavras para análise;
  - Delimitação de um número de palavras para possível análise.
- Análise da padronização das palavras selecionadas por meio de concordâncias, listas de colocados e n-gramas;
- Descrição da padronização das palavras selecionadas como resultado.
- Utilização dos padrões em objetivos seguintes (confecção de atividades, por exemplo).

A pequena metodologia descrita é ponto de partida inicial para uma gama de pesquisas em Linguística de *Corpus*. Os passos podem ser cíclicos e envolver algum tipo de recurso adicional durante o processo.

Ainda que seja uma metodologia básica, exige grande esforço e trabalho do pesquisador, tanto na preparação do material, *corpus*, como na análise, identificação de padrões, por exemplo.

Dado o conhecimento das principais ferramentas e noções metodológicas passaremos a seguir para uma reflexão sobre lacunas e pontos a serem otimizados no processo de pesquisa e possíveis oportunidades para pesquisa e desenvolvimento de recursos da adaptatividade em Linguística de *Corpus*.

VII. O microscópio da Linguística de *Corpus* pode ter aspectos adaptativos

Sem sobra de dúvida, o computador com suas ferramentas é imprescindível para pesquisas em Linguística de *Corpus*, e em muitas outras áreas, o que justifica a citação da metáfora do microscópio para a área. Muitas descobertas sobre a língua têm sido reveladas por meio de sua instrumentação, a ponto de se considerar uma nova maneira de estudar a língua.

Contudo, é preciso sempre fazer uma avaliação dos métodos e recursos utilizados a fim de obter avanços e facilitar o trabalho de pesquisa. É natural que se deseje um microscópio cada vez mais preciso, abrangente e dinâmico, que possa ser utilizado facilmente em diferentes tarefas.

O argumento é o de que muitas das tarefas executadas em pesquisas de Linguística de *Corpus* podem ser beneficiadas com a introdução do conceito de Adaptatividade, principalmente aplicações desenvolvidas para determinada atividade.

Novamente citando a pesquisa de [6], embora houvesse uma série de ferramentas disponíveis, as quais algumas delas já foram descritas, uma nova ferramenta foi criada especificamente para o trabalho de identificação de possíveis candidatos a termo, ZExtractor, disponível em http://www.fflch.usp.br/dl/li/x/?p=559, a qual utilizou funções similares às de concordanciadores e listadores de palavras. O fato de que nem sempre as ferramentas disponíveis serão totalmente adequadas às necessidades da pesquisa é natural e, por isso, há a necessidade de adaptá-las.

No caso, o que se tem feito é tornar as ferramentas adaptáveis, quando possível. Como é a ferramenta criada ZExtractor.

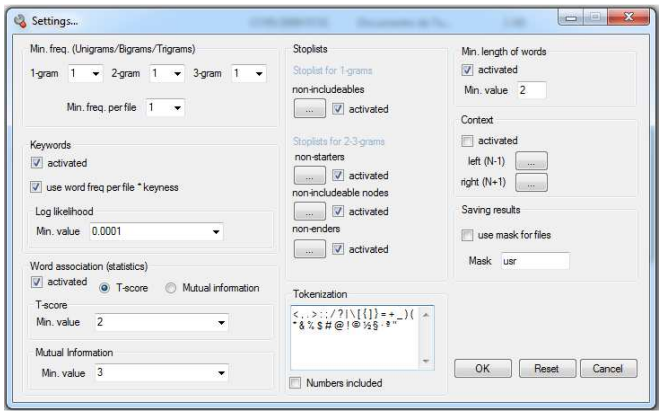

Fig.10. Configurações do Programa ZExtractor

O programa permite que o usuário faça uma série de ajustes para que a execução das funções possa retornar os resultados desejados ou próximos. As opções envolvem a definição de: frequência mínima de ocorrência do candidato no *corpus*, ocorrência mínima do candidato nos arquivos de texto do *corpus*, definição de *stoplists*, valor de corte de chavicidade, frequência do candidato por arquivo e estatísticas de associação de palavras.

Porém, as funções são estáticas e puramente quantitativas, dependendo da calibragem do pesquisador que analisará, a cada saída, os resultados obtidos. Não que seja algo errado, mas se há um modo que possivelmente melhore a execução das tarefas, é preciso investigá-lo, como a própria pesquisa citada o faz, ao testar as várias ferramentas e a partir delas desenvolver uma metodologia para extração de candidatos a termo.

Deste modo, fazemos uma reflexão sobre as ferramentas computacionais e metodologia básica descrita, com o objetivo de apontar alguns problemas enfrentados e possíveis pontos para o desenvolvimento de novas soluções para pesquisa. Ao fazê-lo, tentamos relacionar a pesquisas e trabalhos que utilizam o conceito da Adaptatividade.

VIII. Adaptatividade para pesquisas em Linguística de *Corpus*

Nesta seção, fazemos a descrição de alguns problemas e limitações em pesquisas de Linguística de *Corpus*, entendendo que o conceito de adaptatividade e técnicas adaptativas poderiam ser empregados para criação de soluções dinâmicas. É importante destacar que as descrições fornecidas estão sob a ótica de pesquisas relacionadas ao uso da Linguística de *Corpus* para o desenvolvimento de aplicações, cursos e materiais pedagógicos para o ensino de línguas.

Em relação à coleta e compilação de *corpus*, às vezes não é possível contar com uma grande quantidade de textos para um *corpus*, por várias restrições (direitos autorais, por exemplo), o que pode prejudicar sua utilização em determinada aplicação em que necessite de grandes quantidades de dados para obter um resultado confiável. É o caso de aplicações baseadas em dados estatísticos. O uso de dispositivos baseados em regras poderia auxiliar em uma solução.

Em muitas pesquisas, os objetivos, a metodologia de análise e aplicação podem ser os mesmos, tendo apenas a distinção do *corpus* e/ou gênero, como é o caso de algumas pesquisas que procuram extrair a padronização de palavras e fraseologias para uso pedagógico. Seria desejável o desenvolvimento de uma solução comum que pudesse lidar com as diferenças, sem que o pesquisador tenha que passar por todos os passos novamente. Em termos metodológicos, pode parecer que pesquisadores estejam fazendo sempre o mesmo esforço. O caminho poderia ser encurtado.

Algumas ferramentas computacionais para análise de *corpus* possuem limites de processamento de textos. O Wordsmith tools 3.0, por exemplo, no Concord, limita o número de linhas de concordância retornadas. Se a palavra ocorre mais do que o limite, não é possível analisá-la em sua totalidade, sendo necessário fazer um recorte e estimativas.

O recorte é muito comum nas pesquisas. Pode estar relacionado à quantidade de itens a ser analisada, tempo e esforço para a pesquisa. Contudo, todos esses fatores poderiam ser minimizados, sem que o pesquisador tivesse que executar a mesma pesquisa em dois momentos, como condição o aumento quantitativo dos itens analisados. O desenvolvimento de soluções de treinamento e aprendizado de máquina poderia auxiliar em tarefas que filtrassem itens ou executassem análises com dados previamente armazenados.

O recurso de extração de palavras-chave para filtragem de palavras a serem analisadas, que nas ferramentas aqui descritas são identificadas a partir de fórmulas estatísticas, poderia ser incrementado com regras dinâmicas para a separação, por exemplo, de palavras relacionadas especificamente ao assunto/conteúdo do texto, em relação às palavras comuns ao gênero dos textos do *corpus*, uma necessidade típica da pesquisa em Linguística de *Corpus*.

Retornando ao exemplo do programa ZExtractor, ao invés de especificar regras/configurações comuns para todas as palavras do corpus, o conjunto de regras poderia ser aplicado dinamicamente para cada possível candidato, podendo em alguns casos até solicitar a intervenção do usuário e memorizar sua ação para aplicações futuras.

O reconhecimento de padrões, uma das tarefas mais importantes de pesquisas em Linguística de *Corpus* poderia ser beneficiado por soluções adaptativas, uma vez que na área da Tecnologia Adaptativa há vários trabalhos nesse sentido.

Como já é de conhecimento, o computador lida melhor que o ser humano em tarefas repetitivas, em que o nível de atenção deve ser alto e contínuo. Descobertas interessantes em análises poderiam surgir.

A partir dessas poucas considerações, é possível afirmar que estudos que busquem adição do conceito e técnicas adaptativas em pesquisas da Linguística de *Corpus* poderiam transformar a maneira em que seu microscópio é utilizado.

Na seção a seguir, citamos aspectos de uma pesquisa em andamento que apresenta passos iniciais para o uso da Adaptatividade no desenvolvimento de um sistema de montagem automática de atividades de leitura em língua inglesa com *corpora*. O objetivo é salientar as possíveis contribuições da Tecnologia Adaptativa para a área de Linguística de *Corpus*, especificamente em relação à linha de pesquisa que se preocupa com o desenvolvimento de aplicativos e ferramentas para análise de corpora e ensino de línguas.

IX. Gerador de Atividades de Leitura Adaptativo

O desenvolvimento de um sistema de geração e montagem de atividades de leitura tem como objetivo prático de suprir a necessidade de professores que desejam utilizar materiais baseados em *corpora* em suas aulas, 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.

O projeto está baseado em um estudo realizado em uma pesquisa de mestrado [7], que teve como produto final um *software desktop* para preparação semiautomática de atividades de leitura em inglês, tomando como entrada um texto selecionado pelo usuário com fins pedagógicos e conduzindo-o, através de etapas, como um assistente eletrônico, até a publicação de uma unidade didática com exercícios baseados em concordâncias (*data-driven learning*), predição, léxico-gramática e questões para leitura crítica, utilizando análises automáticas do texto selecionado por meio de fórmulas estatísticas: lista de frequência, palavras-chave, possíveis palavras cognatas, etiquetagem morfológica, possíveis padrões (*clusters*) e densidade lexical do texto.

Nesse primeiro protótipo, com base no conceito de '*standard exercise*' (atividade padrão) introduzido em [8] para cursos de ESP (*English For Specific Purposes*), 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.

Tendo em vista tais necessidades, a atual pesquisa estuda a possibilidade do uso da Tecnologia Adaptativa para o desenvolvimento de uma aplicação dinâmica, em que o conjunto de exercícios seja variado, flexível e adequado à entrada (texto/gênero). Em contraponto a um modelo padrão de atividade, como no primeiro protótipo, propõe-se o conceito de um gerador adaptativo de exercícios.

O esquema representado a seguir ilustra as etapas básicas para o funcionamento do sistema para que seja possível a inclusão de técnicas adaptativas na geração automática de exercícios de leitura, tendo como partida um texto e opções pré-definidas de um usuário.

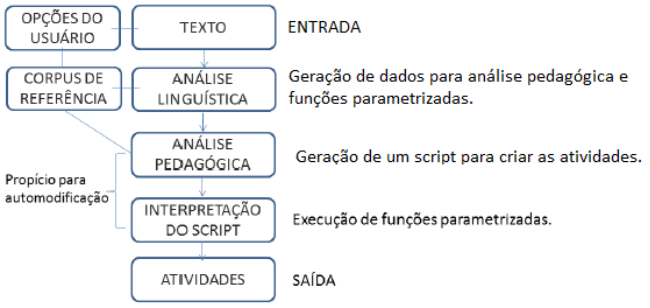

Fig. 11. Esboço para uma aplicação dinâmica

A implementação do sistema será feita utilizando a linguagem Python. Inicialmente, criamos um módulo para análise linguística de texto e corpora. Os dados da análise servirão posteriormente para tomadas de decisão em um módulo de análise pedagógica do texto, que auxiliará na composição de um script em xml, conforme a seguir:

```
<?xml version="1.0" ?>
<activity>
<label content='Qual conjunto de palavras possui substantivos?\n' />
<label content= '\na) ' /> <pos tag='NN' limit='5' random='true' order='none' delimiter=' - '
case='upper' length='5'  />
<label content= '\nb) ' /> <pos tag='JJ' limit='5' random='true' order='none' delimiter='; '
case='upper' />
<label content= '\nc) ' /> <pos tag='VB|JJ' limit='5' random='true' order='none' delimiter=' '
case='upper' />
<label content= '\nd) ' /> <pos tag='VB' limit='5'  />
<label content='\n\n' />

<label content='Qual das alternativas possui o maior numero de palavras cognatas?\n' />
<label content= '\na) ' /><cognates limit='8' random='false' order='none' case='lower' length=
'0' delimiter='; ' />
<label content= '\nb) ' /><keywords limit='8' random='false' order='desc' case='lower' length=
'0' stoplist='true'  delimiter='; ' />
<label content= '\nc) ' /><keywords limit='8' random='false' order='asc' case='lower' length='0'
 stoplist='true'  delimiter='; ' />
<label content= '\nd) ' /><keywords limit='8' random='true' order='asc' case='lower' length='4'
stoplist='true'  delimiter='; ' />
</activity>
```

Fig. 12. Exemplo de script para criação de exercícios

Após a montagem do script, sua interpretação é realizada por meio de funções parametrizadas. O resultado final é a impressão de exercícios para as atividades:

```
Problems  Console
<terminated> C:\Users\home\Documents\workspace\TEXT\test_publish.py
Qual conjunto de palavras possui substantivos?

a) FLOES - COMPANIES - IMPLICATIONS - SCIENTISTS - DEVELOPMENT -
b) NEW; ACIDIC; TALLER; LAND-BASED; DARK;
c) FOUND LIKELY REPORTED WARMER OPEN
d) sent warmed affect said bears

Qual das alternativas possui o maior numero de palavras cognatas?

a) global; including; previous; percent; annual; absorbs; international; report;
b) arctic; ice; sea; report; normal; scientists; polar; u.s.;
c) summer; taller; clog; mingle; help; corps; increasingly; when;
d) setting; national; carbon; research; never; 2007,; until; at;
```

Fig. 13. Interpretação de script para criação de exercícios

As descrições do sistema em questão caminham para o entendimento e aplicação de técnicas adaptativas. A montagem dinâmica do script para a criação dos exercícios, com base nas análises linguística e pedagógica, considerando as opções definidas pelo usuário, pode conferir aspectos adaptativos ao sistema, o que elevariam o conceito de 'atividade padrão' a um conceito de 'atividade adaptativa' para elaboração de materiais pedagógicos para o ensino de leitura em língua estrangeira.

Embora haja muito trabalho a ser realizado, a investigação é importante para o estabelecimento de um diálogo concreto entre Linguística de *Corpus* e Tecnologia Adaptativa, especificamente no desenvolvimento de aplicações dinâmicas para análise de corpora e ensino de línguas.

X. Considerações Finais

O trabalho buscou mostrar as principais práticas e necessidades em pesquisas de Linguística de *Corpus*, em uma linha de desenvolvimento de ferramentas para análise de *corpus* e ensino de línguas, abordando questões de uso de ferramentas computacionais e metodológicas, as quais poderiam ser foco de investigação em trabalhos futuros com o diálogo entre áreas da Linguística de Tecnologia Adaptativa.

Podemos considerar as informações apresentadas como um primeiro esboço para tal diálogo. Esperamos que seja útil para fomentar uma série de discussões sobre o assunto.

REFERÊNCIAS

[1] Guariento W. & Morley J. (2001). Text and task authenticity in the EFL classroom. ELT Journal. 55/4 October, 1-7.

[2] GILMORE, A. (2004). A comparison of textbooks and authentic interactions. ELT Journal,58/4 October, 1-12.

[3] DIZERÓ, W. Formalismos Adaptativos Aplicados na Modelagem de *Software*s Educacionais. Tese de Doutorado, EPUSP, São Paulo, 2010.

[4] PADOVANI, D.; CONTIER, A.; JOSÉ NETO, J. J.; Tecnologia Adaptativa Aplicada ao Processamento da Linguagem Natural. In: WTA 2010; Quarto Workshop de Tecnologia Adaptativa, 2010, São Paulo. Memórias do WTA 2010: Quarto Workshop de Tecnologia Adaptativa. São Paulo: Laboratório de Linguagens e Técnicas Adaptativas, 2010. p. 35-42.

[5] BERBER SARDINHA, T. (2004) *Linguística de Corpus*. São Paulo: Manole.

[6] SILVA E TEIXEIRA, R.B. Termos de (Onco)mastologia: uma abordagem mediada por corpus. 2010. Dissertação de mestrado. Pontifícia Universidade Católica de São Paulo.

[7] MOREIRA FILHO, P. Desenvolvimento de um software para preparação semiautomática de atividades de leitura em inglês. Dissertação de Mestrado Inédita, LAEL, PUC-SP, 2007.

[8] Disponível em: < http://wordsmithtools.com/downloads/corpus_linguistics/Standard%20Exercise.pdf>. Acesso em 01/02/2013.

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 de Centro de Línguas da Diretoria Regional de Ensino 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.