Adaptatividade em um sistema de criação de atividades de leitura para o ensino de línguas

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

***Resumo* — Neste trabalho, apresenta-se a descrição de um estudo sobre o uso de abordagens de análise de corpora da Linguística de Corpus, aliadas a técnicas de Processamento de Língua Natural (PLN), Aprendizagem de Máquina (AM) e Tecnologia Adaptativa (TA), para o desenvolvimento de um protótipo de sistema de processamento de língua natural capaz de gerar, automaticamente, atividades de leitura em língua inglesa a partir de um texto e corpus. A proposta do sistema concentra-se na automatização das análises linguísticas, para a criação automática de atividades de leitura. Assim, são apresentados os objetivos e premissas do estudo, a metodologia, uma pequena descrição do processo de desenvolvimento do sistema e os avanços em sua arquitetura para a implementação da adaptatividade.**

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

I. INTRODUÇÃO

Este trabalho apresenta os avanços obtidos no trabalho de pesquisa na criação de um sistema de processamento de língua natural para a geração de atividades de leitura em língua inglesa com corpora.

Embora o uso de materiais baseados em corpus seja positivo para o processo de ensino-aprendizagem, a preparação e elaboração desses materiais ainda não é uma realidade 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 para que possa aproveitar os benefícios do uso de corpus e suas ferramentas computacionais de análise 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 PLN, de práticas de análise de corpus para o ensino de línguas, esta pesquisa tem como um de seus objetivos tentar suprir a necessidade de professores de língua estrangeira 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.

A investigação está baseada em um estudo inicial realizado em uma pesquisa de mestrado [2], que teve como produto final um *software desktop* para preparação semiautomática de atividades de leitura em inglês. O produto final do estudo considera como entrada um texto selecionado pelo usuário e, através de etapas, como um assistente eletrônico, levá-lo até a publicação de uma unidade didática.

Nos primeiros protótipos, com base no conceito de *standard exercise*[1] (SCOTT et al., 1984, p. 1) para o ensino de leitura de *English For Especific Purposes* (*ESP*)[2], um conjunto fixo de exercícios é preparado automaticamente, incluindo atividades baseadas em concordâncias, *data-driven learning*[3], 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 a consolidação de uma ferramenta que possa ser usada pelo usuário final. Há situações em que o programa gera exercícios com erros, como incluir uma palavra de conteúdo na lista de palavras gramaticais. Nesses casos, há a possibilidade de intervenção do usuário, mas, se a quantidade de erros for numerosa de forma a exigir constantemente esse tipo de intervenção do usuário/professor, o sistema pode perder sua utilidade, impossibilitando seu uso pedagógico.

Outro ponto essencial é o aumento da variedade de exercícios disponíveis e a quebra da limitação do usuário a um modelo fixo. O conceito de atividade padrão está relacionado à necessidade de treinar a compreensão leitora do aprendiz a partir de um conjunto fixo de questões que

[1] atividade padrão, tradução nossa.
[2] inglês para fins específicos.
[3] ensino movido a dados, tradução nossa.

poderiam ser utilizadas a quase qualquer texto. A invariabilidade do modelo padrão faz com que o programa tenha sua funcionalidade limitada e estática. A adição de recursos adicionais significaria toda a sua reprogramação.

Em uma nova proposta de sistema, a geração de exercícios em relação a outros itens comuns no ensino de estratégias de leitura e léxico-gramática, adequados a uma gama maior de textos é pretendida. Em vez de um modelo fixo e estático, procura-se desenvolver um ambiente propício com um conjunto de instruções, parâmetros e regras que possibilitem uma funcionalidade flexível e dinâmica, em que a criação e montagem dos exercícios sejam feitas autonomamente de acordo com o texto de entrada, sem a necessidade de reprogramação.

Desse modo, buscando estender os objetivos da pesquisa de mestrado [2], empregam-se, em diálogo, pressupostos teóricos e metodológicos, além de técnicas das áreas de Linguística de Corpus, Processamento de Língua Natural, Aprendizagem de Máquina e Adaptatividade em todos os níveis aplicáveis de um sistema de processamento de língua natural, que, a partir do fornecimento de um texto e/ou algum tipo de entrada preestabelecida, gere atividades didáticas de leitura em língua inglesa.

A princípio, considera-se relevante a exploração do trabalho em quatro níveis/camadas do sistema: i. Opções do usuário; ii. Análise linguística da entrada; iii. Análise pedagógica; iv. Montagem de exercícios. A atividade de ensino a ser gerada deverá levar em consideração as informações de todos os níveis. O produto final estará condicionado às opções do usuário (seleção de itens de ensino, tipos de exercício e itens léxico-gramaticais incluídos), análise linguística de textos (informações de frequência, palavras-chave, anotação morfossintática, entre outras), análise pedagógica da entrada (a partir da análise linguística, quais tipos de exercícios são possíveis e adequados), montagem de exercícios (como extração e organização de itens linguísticos).

Embora ainda haja muito trabalho a ser realizado, a investigação é importante para o estabelecimento de um diálogo concreto entre Linguística de Corpus, Processamento de Língua Natural, Aprendizagem de Máquina e Adaptatividade, especificamente no que diz respeito ao desenvolvimento de aplicações dinâmicas para análise de corpora e ensino de línguas.

II. OS OBJETIVOS E PREMISSAS DO ESTUDO

O objetivo geral desta pesquisa é conceber um sistema de processamento de língua natural capaz de analisar corpora e textos para gerar, automaticamente, informações pedagógicas sobre o texto, atividades de leitura e ensino de padrões (léxico gramática) em língua inglesa.

Os objetivos específicos para consecução do objetivo geral envolvem:

- estudar as abordagens e ferramentas de análise de corpora para a criação de materiais de ensino na área de Linguística de Corpus;
- estudar as abordagens, técnicas, algoritmos e recursos das áreas da computação como Processamento de Língua Natural, Aprendizado de Máquina e Adaptatividade e suas tecnologias;
- estudar a implementação de algoritmos por meio de programação em linguagem Python;
- coletar e analisar a padronização de corpora em gêneros específicos a fim de serem utilizados no desenvolvimento e avaliação de um sistema de análise de corpora e textos;
- construir módulos de análise de corpora e textos para obtenção de recursos e dados linguísticos para avaliação pedagógica de um texto e criação de exercícios;
- avaliar por meio de experimentos práticos os módulos de análises construídos, além de sua eficácia na criação de atividades;
- contrastar os benefícios de uso do sistema em relação às ferramentas de programas de análise de corpora, especificamente os concordanciadores;
- formalizar a arquitetura do sistema;
- apresentar uma análise de problemas com soluções adotadas;
- apresentar possíveis adaptações do sistema para uso de textos em língua portuguesa.

As principais premissas da pesquisa são as de que:

- as ferramentas de análise de corpora atuais, por serem desenhadas para uso geral, em diferentes propósitos, podem demandar grande esforço de adaptação de suas funcionalidades para a pesquisa e construção de materiais baseados em dados de análise de corpora;
- a preparação de materiais no ensino de línguas pode ser enriquecida com o auxílio de recursos computacionais que contribuam para a otimização dos estudos linguísticos em larga escala, tanto em relação ao tempo e esforço, quanto em qualidade de seus resultados;
- a automatização de análises e exploração de corpora para fins específicos, como a extração de dados para elaboração de materiais didáticos e atividades para o ensino de línguas, pode ser uma realidade, a partir do estudo das ferramentas computacionais disponíveis e o diálogo entre outras áreas de estudo;
- os dados de análises básicas, comuns no processamento de língua natural, podem ser extrapolados na criação de atividades de leitura e ensino de léxico-gramática;
- o uso de uma abordagem específica de análise automática de corpus pode trazer benefícios significativos para a tarefa de elaboração de materiais de ensino baseados em corpora;

um maior diálogo da Linguística de Corpus com outras áreas que também fazem uso de corpora, as quais também fazem uso de corpora, especificamente Processamento de Língua Natural, Aprendizagem de Máquina e

Adaptatividade, pode otimizar os procedimentos de pesquisa e análise de corpora;

a formalização da necessidade de análise automática de corpora pode ser objeto interessante de pesquisa para o linguista de corpus em novos caminhos de investigação.

III. METODOLOGIA

A abordagem de desenvolvimento do sistema de análise de corpora e textos em língua inglesa para a criação automática de atividades de leitura e ensino de padrões baseia-se nas funcionalidades das ferramentas existentes de análise e exploração de corpora, como ponto de partida.

A metodologia do processo de desenvolvimento consiste em:

- analisar os elementos constituintes das principais ferramentas de análise e exploração de corpora e as metodologias de uso de tais ferramentas no âmbito do ensino de línguas relacionado à Linguística de Corpus;
- coletar e compilar os corpora de estudo, desenvolvimento e teste para todo o processo de desenvolvimento do sistema proposto;
- estudar, construir e adaptar métodos, técnicas e algoritmos existentes para análise de corpora e textos, necessários para as análises pretendidas;
- anotar dados de corpora para o treinamento e teste de algoritmos;
- elaborar módulos de análise de apoio, por exemplo, módulos de itemização, etiquetagem, contagem de palavras, extração de palavras-chave, extração de n-gramas entre outros, para a análise de corpora e textos;
- aplicar os módulos de análise nos corpora de estudo, desenvolvimento e teste, a fim de obter a descrição dos corpora e possibilidades de dados e recursos a serem utilizados para a criação de atividades;
- avaliar e otimizar os módulos de análise a partir de sua aplicação nos corpora de desenvolvimento e teste;
- explorar, criar e adaptar os módulos de análise na construção de algoritmos de criação de atividades de ensino.

IV. DESENVOLVIMENTO

Para o desenvolvimento da pesquisa, utilizamos a linguagem de programação Python e a biblioteca de processamento de língua natural NLTK para a criação dos módulos de análise. A implementação da interface foi realizada as linguagens PHP, HTML, CSS e javascript.

Para o desenvolvimento da pesquisa, utilizamos dois principais tipos de corpora:corpus de referência e corpus de treinamento/estudo. Os corpora de referência são utilizados para comparação entre os corpora de treinamento e estudo, além de servirem para a extração de padrões léxico-gramaticais. Os corpora de treinamento/estudo são utilizados para o desenvolvimento das funções de análise do sistema e para a realização de testes.

Os principais corpora do estudo são:

- Corpus de língua geral do inglês;
- Corpus de anúncios de emprego em inglês;
- Corpus de artigos em inglês da revista *Scientific American*.

Além dos corpora descritos, foram utilizados os recursos e léxicos da biblioteca *NLTK* (*Natural Language Tool Kit*), tal como o corpus *Penn Treebank*.

As principais etapas de desenvolvimento do sistema foram a criação de:

- módulos de análise linguística de texto em linguagem Python;
- um módulo que compila todas as análises e cria um arquivo do texto analisado em *XML*;
- um módulo de leitura e extração de informações de um texto analisado em *XML*;
- um módulo de leitura e extração de informações de um corpus formado por textos já analisados em *XML*;
- um módulo interpretador que recebe como entrada um código *XML (templates)* e retorna, a partir de funções parametrizadas, uma atividade ou exercício;
- Criação da interface com o usuário.

Embora os módulos sejam projetados para análise de textos em língua inglesa, sua criação foi desenhada para permitir a análise de textos de outras línguas.

Para cada módulo, há uma opção de seleção de língua. O uso do módulo para uma outra língua pode ser conseguido, acrescentando-se um novo conjunto de regras, léxico ou outro tipo de dado, conforme a especificidade da análise.

Por exemplo, no módulo de extração de palavras-chave, há o diretório '*data/en*' e, dentro, um arquivo de referência '*reflist.txt*'. Para extração de palavras-chave em português, bastaria adicionar um novo diretório e arquivo de referência ('*data/br/reflist.txt*') e, no código, na instanciação da classe de análise, passar o argumento *'br*' como língua de análise.

Os principais módulos de análise linguística compreendem: segmentação, itemização, etiquetagem morfossintática, contagem de palavras, extração de palavras-chave, extração de n-gramas, identificação de palavras-cognatas, identificação de afixos, identificação de grupos nominais simples (*noun phrases*), identificação de referência pronominal, reconhecimento de entidades nomeadas e extração de relações. Tendo em vista a arquitetura projetada, é possível estender as a qualquer momento, com a adição de novos módulos.

As análises podem propiciar uma série de informações para a análise pedagógica do texto e matéria prima para atividades de leitura e ensino de lexicogramática, que podem, a princípio, serem geradas em formato de questão aberta ou formato de múltipla escolha, havendo a possibilidade de já indicar a resposta correta. Como exemplo ilustrativo, a análise automática pode identificar no corpus de anúncios de empregos as entidades '*COMPANY*' e '*POSITION*' (marcadas com códigos) na seguinte sentença:

*<**ne type='COMPANY'**>WCMC<**/ne**> are now seeking a permanent <**ne type='POSITION'**>Head of Programme<**/ne**> to work with us.*

A partir da análise, é possível criar questões sobre qual é o nome da empresa ou qual vaga. A identificação de tais entidades podem também contribuir para a análise dos movimentos ou passos do gênero. Pode-se criar atividades de localização de informações, a partir de perguntas como: em qual parte do texto podemos encontrar a vaga disponibilizada.

Outro exemplo é a análise quantitativa realizada das palavras do texto. Pode-se criar uma atividade em que deve-se identificar as palavras mais frequentes do texto e/ou quais são suas palavras-chave. A intenção é aproveitar ao máximo as análises para os fins pedagógicos de leitura e ensino de léxicogramática.

A programação dos módulos para as análises como as descritas foi realizada, em grande parte, a partir de heurísticas, métodos e algoritmos utilizados nas áreas de Linguística de Corpus, Processamento de Língua Natural e Aprendizagem de Máquina. Tendo em vista o número de análises e a complexidade do processamento de língua natural, há limitações reconhecidas que poderão minimizadas conforme a evolução do sistema, seja por meio de acréscimos de novas regras ou aumentos dos léxicos e dados de treinamento.

V. TECNOLOGIA ADAPTATIVA

Conforme [4], 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 automodificar 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 [5]: 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.).

Tendo em vista a complexidade do estudo pretendido, uma das áreas de interesse para a pesquisa, em relação à Tecnologia Adaptativa, é o processamento de linguagens naturais. A aplicação da Tecnologia Adaptativa ao processamento de línguas naturais é um campo de extrema importância.

VI. ARQUITETURA DO SISTEMA

O sistema recebe um texto como entrada e tem como objetivo realizar análises linguísticas automáticas que permitem sua exploração para extração de informações relevantes para a análise pedagógica do usuário e a criação automática de atividades.

Descrevemos aqui a arquitetura planejada para a implementação da adaptatividade no sistema, tendo em vista que sua interface está ainda em construção. A figura a seguir ilustra a arquitetura do sistema.

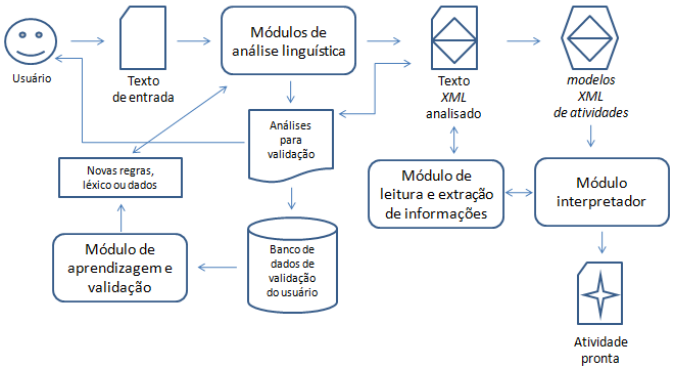

Figura 1. Arquitetura do sistema

A arquitetura apresentada pode ser descrita a partir dos seguintes passos:

1. O usuário insere o texto de entrada, o qual será o foco para a criação das atividades de leitura;
2. Os módulos de análise linguística processam o texto e geram um conjunto de itens de análise automática para validação do usuário;
3. O usuário valida e/ou corrigi os itens de análise automática;
4. Os itens validados e/ou corrigidos são armazenados em um banco de dados;
5. Um módulo de aprendizagem e validação analisa o banco de dados de validação do usuário e gera novas regras, léxicos e/ou dados de aprendizagem;
6. Os módulos de análise linguística são atualizados com os dados gerados pelo módulo de aprendizagem e validação;
7. Os módulos de análise linguística geram um arquivo *XML* com a análise do texto de entrada;
8. Modelos de atividades escritos em XML são instanciados para o texto analisado (selecionados pelo usuário ou requeridos a partir das características do texto analisado);

9. Um interpretador traduz cada instrução dos modelos de atividade e, por meio de um módulo de leitura e extração de informações, os modelos são preenchidos para gerar as atividades;
10. O sistema retorna a atividade pronta.

Embora não especificado em detalhes no esquema da figura 1., os modelos de atividades escritos em *XML* poderiam ser gerados no momento de execução conforme as características do texto analisado. Tal funcionalidade estaria relacionada a uma análise pedagógica automática, em que seriam determinados quais tipos de exercícios e itens seriam mais adequados dada a análise do texto de entrada.

Entende-se que um texto pode ser explorado de diversas formas. Por exemplo, um professor pode preferir explorar as características gramaticais do texto em relação a sua organização discursiva. Neste caso, modelos de atividades gramaticais seriam instanciados preferencialmente em relação a outros, conforme as características do texto de entrada.

Uma forma de implementação seria a criação de perfis, conforme as possibilidades de exploração pedagógica de um texto. A figura a seguir ilustra o processo de seleção de modelos de atividades:

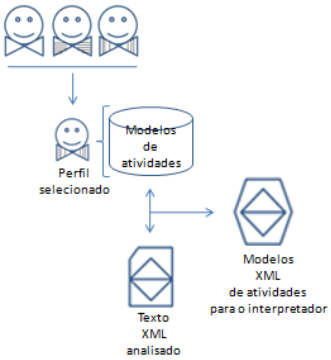

Figura 2. Arquitetura do sistema

Para cada perfil, um conjunto de modelos de atividades estaria disponível. A seleção de um determinado perfil implicaria na instanciação de determinados modelos de atividades que seriam adicionados de acordo com regras definidas, cujos argumentos seriam satisfeitos com base nas análises automáticas do texto de entrada. Após as adições dos modelos de atividades, todo o conjunto é submetido ao interpretador para gerar as atividades.

Como já citado, o sistema está em constante modificação. Os módulos de análise em sua maioria estão consolidados. Os passos seguintes incluem a construção dos módulos necessários para a validação e correção das análise automáticas, realizadas pelo usuário no momento de entrada do texto, e o gerenciamento da aprendizagem a partir do banco de dados das ações do usuário. Considera-se que tais modificações contribuam para conferir uma característica adaptativa ao sistema, o que promoverá sua constante evolução e aumento de suas funcionalidades para a criação de atividades.

VII. Considerações finais

O trabalho buscou mostrar alguns dos avanços feitos em uma pesquisa no desenvolvimento de um sistema de processamento de língua natural 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.

Os resultados da investigação contribuem para fomentar o interesse em relação ao desenvolvimento de ferramentas computacionais para exploração de corpora na automatização da extração de dados linguísticos para a elaboração de materiais de ensino e currículo.

O percurso do desenvolvimento do sistema é considerado parte integrante da pesquisa, uma vez que seu cenário pode ser considerado pouco usual, dada a natureza interdisciplinar do estudo e suas exigências, tendo o pesquisador linguista de enveredar pela análise computacional por meio de programação.

O esforço pode propiciar um contato maior com as diferentes áreas e trazer benefícios para a construção do diálogo pretendido, não só nesta pesquisa, mas também nas pesquisas de linguagem via computador em geral, em que há separação entre equipes: de um lado, aqueles que fazem a programação para a realização de tarefas automaticamente; de outro, aqueles que apenas fazem uso dos instrumentos.

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] 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.

[3] SCOTT, M., CARIONI, L., ZANATTA, M., BAYER, E. &QUINTANILHA, T. Using a Standard Exercise in Teaching Reading Comprehension.English Language Teaching Journal, v. 38, n. 2, pp. 114-20.

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

[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.