

Uso da Tecnologia Adaptativa em um SPLN para criação automática de atividades de leitura

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

Abstract— This paper presents the initial steps of a research on the use of the Adaptative Technology to develop a natural language processing system for creating reading activities in English. More specifically, it proposes the application of adaptative devices in pronominal anaphora resolution for textual cohesion exercises.

Keywords— Natural Language Processing, Corpus Linguistics, Adaptative Technology, Anaphor Resolution.

I. INTRODUÇÃO

ATUALMENTE, é crescente o número de pesquisas sobre o emprego de materiais de ensino baseados em *corpora*, uma vez que privilegiam a língua em uso.

Embora seja desejável o uso de materiais baseados em *corpus* e haja uma série de ferramentas computacionais que auxiliam na análise de *corpora* para a sua criação, seu processo de preparação ainda é muito demorado e, geralmente, realizado apenas por pesquisadores; muitas vezes, requer a análise prévia de grandes quantidades de dados, 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.

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.

Dentre as ferramentas computacionais que podem auxiliar na análise de *corpora*, citamos dois programas de computador desenvolvidos especialmente para o trabalho com textos e que reúnem recursos computacionais, matemáticos e estatísticos – *WordSmith Tools* e *Stablex PC*.

De autoria de Mike Scott, Universidade de Liverpool, o programa *WordSmith Tools* (1998) é publicado pela Oxford University Press e distribuído via *World Wide Web* (<http://www.lexically.net/wordsmith>). Em sua quinta versão, disponível para PC/Windows 2000 ou superior, incluindo Windows 7, disponibiliza diversos recursos – *Wordlist*, *Concord*, *Keywords*, *Splitter*, *Text Converter*, *Dual Text Aligner*, *Viewer* – para tarefas específicas de análises de textos.

O programa *Stablex* (STA – de statistique, TAB – de tableaux, LEX – de lexique e T...EX – de texte), de autoria de André Camlong e Thierry Beltran, Universidade de Toulouse II, inicialmente desenvolvido para *Macintosh* (Toulouse, Teknea, 1991), conta, atualmente, com a sua versão PC (São

Paulo, Pirus Tecnologia, 2004). Os seus recursos – geração de léxicos, indexação, extração de sequências e concordâncias, lematização, tratamento estatístico – foram desenvolvidos em função de um modelo de análise lexical, textual e discursiva – *método matemático-estatístico-computacional de análise de textos* de André Camlong. Trata-se, por conseguinte, da aplicação de um programa que serve de ferramenta para um método de análise de textos.

O método é fundado na matemática e na estatística paramétrica (estatística descritiva); possibilita o estudo descritivo, objetivo e indutivo do texto; permite a análise quantificativa do léxico, que indica apontamentos para a análise textual e discursiva. Nele, o texto é o ponto de referência: as operações estatísticas partem do texto e, por sua vez, refletem o texto.

Destacamos o fato de as listas originárias da análise de textos pelo *Stablex* exibirem o léxico do *corpus* e dos textos que o integram – e não apenas do *corpus* como um todo, caso do *WordSmith* –, o que facilita uma visão contrastiva do todo – *corpus* – em relação às partes – textos que integram o *corpus* – e das partes em relação ao todo, bem como das partes entre si. Já o *WordSmith* assume importância fundamental no tratamento de *corpora* extensos, voltados à construção de dicionários e de glossários, aos exames dos padrões linguísticos, aos estudos ligados à tradução e ao gênero.

A partir do desenvolvimento, aplicação e análise de um sistema de montagem automática de atividades *online* de leitura em língua inglesa com *corpora*, por meio do uso de técnicas adaptativas, a pesquisa em andamento tem o objetivo 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.

A investigação está baseada em um estudo realizado em uma pesquisa de mestrado [2], 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. A versão mais recente do *software*, *Reading Class Builder*, está disponível no sítio: <http://www.fflch.usp.br/dl/li/x/?p=409>.

Nos primeiros protótipos, diferentes exercícios são preparados automaticamente, incluindo atividades baseadas em concordâncias (*data-driven learning*), 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-

¹ Textos autênticos, não inventados para ensinar língua.

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.

Há situações em que o programa gera exercícios com erros, tal como incluir na lista de palavras gramaticais uma palavra de conteúdo. Nesses casos, o usuário pode fazer a correção manualmente na atividade gerada. Porém, se a quantidade de erros for numerosa e exigir constantemente esse tipo de intervenção do usuário/professor, o sistema pode perder sua utilidade, impossibilitando seu uso pedagógico.

Outro ponto importante a ser foco de melhorias é a pouca variedade de exercícios disponíveis e a limitação do usuário a um modelo fixo. O programa poderia gerar exercícios em relação a outros itens de ensino comuns no ensino de estratégias de leitura, como grupos nominais, marcadores discursivos, grau de adjetivos e advérbios, questões de *skimming* e *scanning*, formas verbais e referência pronominal. Seria desejável também que a interface com o usuário permitisse a seleção de determinados pontos a serem ensinados e que o sistema auxiliasse na escolha de texto e exercícios adequados tais fins.

Neste artigo, descrevemos as possibilidades de uso e aplicação de dispositivos adaptativos em um dos módulos do sistema, cuja função específica é a criação de exercícios baseados em coesão textual. Tal uso será considerado como um estudo piloto em relação à aplicação da Tecnologia Adaptativa.

II. OBJETIVOS

Os principais objetivos do estudo são: i. Apresentar, de maneira sucinta, alguns fundamentos linguísticos em relação à coesão textual; ii. Fazer uma revisão dos principais métodos tradicionais de resolução de anáfora; iii. Propor uma abordagem alternativa pelo uso de dispositivos adaptativos.

A intenção não é realizar uma pesquisa completa em relação à resolução de anáforas pronominais, mas fazer um ensaio de possíveis modelos de pesquisa para adaptação de métodos e abordagens existentes, utilizando conceitos e formalismos da Tecnologia Adaptativa.

Buscando estender os objetivos do mestrado [2], pretendemos empregar a Tecnologia Adaptativa em todos os níveis aplicáveis de um sistema de processamento de língua natural a ser construído, o qual, a partir do fornecimento de um texto e/ou algum tipo de entrada preestabelecida, deverá gerar atividades didáticas de leitura em língua inglesa.

No processo, estudaremos a possibilidade do uso da Tecnologia Adaptativa, a princípio, em quatro níveis/camadas do sistema: i. Opções do usuário; ii. Análise linguística da entrada; iii. Análise pedagógica; iv. Montagem de exercícios.

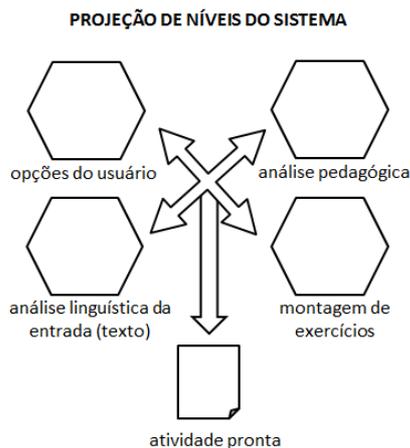


Figura 1. Níveis/camadas do sistema a ser construído

A atividade de ensino a ser gerada deverá levar em consideração as informações de todos os níveis. Desse modo, 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 morfosintática, entre outras), análise pedagógica da entrada (a partir da análise linguística, quais tipos de exercícios são possíveis e adequados), montagem de exercícios (como extrair e organizar os itens linguísticos).

III. PRESSUPOSTOS TEÓRICO-METODOLÓGICOS

O projeto encontra suporte teórico na Linguística de Corpus, Linguística Textual, Linguística Aplicada e Tecnologia Adaptativa.

A Linguística de Corpus, área que estuda os fatos da língua em uso e utiliza o computador no armazenamento e no tratamento e análise de dados empíricos, preocupa-se com a construção de *corpora* eletrônicos a partir de textos e discursos reais.

A Linguística Textual é a ciência que se dedica a investigar a estrutura e o funcionamento dos textos, sendo uma de suas tônicas os estudos sobre coesão e coerência textuais.

A Linguística Aplicada, em sua concepção atual, que vai além da aplicação da Linguística ao ensino/aprendizagem de línguas, dedica-se às situações de uso da língua e, pois, ao desenvolvimento de pesquisas a partir da análise de *corpora*.

A Tecnologia Adaptativa é o principal aporte teórico e metodológico para o desenvolvimento do sistema proposto.

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 auxílio 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 [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.).

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.

IV. FUNDAMENTOS LINGÜÍSTICOS DA COESÃO TEXTUAL

Para Halliday & Hasan [5], a coesão textual é definida como um conceito semântico que se refere às relações de sentido existentes no interior do texto e que o definem como um texto, ocorrendo quando a interpretação de algum elemento no discurso é dependente de outra.

Conforme os autores em [5], cinco principais mecanismos ou fatores de coesão podem ser identificados: referência, substituição, elipse, conjunção e coesão lexical. As relações de referência se desdobram em dois tipos: situacional (exófora) e textual (endófora). Chamamos de referência exofórica aquela em que a referência é feita a algum elemento da situação de comunicação, estando fora dos domínios do texto, sendo a interpretação apenas possível pela análise de tal situação. Na referência endofórica, ao contrário, é possível encontrar a interpretação dentro do próprio texto.

Em Processamento de Língua Natural, é comum a manipulação computacional de elementos de referência endofóricos. As referências endofóricas são divididas em dois tipos de acordo com a posição do referente: anafórica e catafórica. Quando a entidade referenciada antecede o elemento coesivo, temos uma relação anafórica. Quando a entidade referenciada está depois do item coesivo, temos uma relação catafórica.

No exemplo a seguir, destacamos a relação anafórica em que o pronome de terceira pessoa 'ela' faz referência à entidade 'presidenta Dilma Rousseff'.

“Os acordos anunciados durante a visita da presidenta Dilma Rousseff à China significaram um começo razoável no

que ela propôs como um novo capítulo no relacionamento bilateral” [6].

Neste trabalho, tratamos apenas de métodos e abordagens computacionais de resolução de anáforas.

Existem várias definições de anáfora. A palavra anáfora possui origem no grego antigo, composta por ana- 'para trás' e -phora 'ação de levar/transportar'.

Conforme [7], é comum o uso da definição clássica de Halliday e Hasan baseada na noção de coesão. Segundo esses estudiosos, a anáfora é coesão (pressuposição) que aponta de volta para algum item anterior. De acordo com tal definição, a entidade/item no fenômeno de referência que aponta para trás, para algum antecedente, é chamada de anáfora.

Existem muitos tipos de anáfora. Contudo, na literatura de Linguística Computacional, encontramos uma grande quantidade de trabalhos que privilegiam anáforas pronominais (*pronominal anaphoras*), que são realizadas por pronomes anafóricos, como os de terceira pessoa.

Há também uma distinção entre anáforas intrassentenciais e interssentenciais. Anáforas intrassentenciais referem-se aos antecedentes na mesma sentença. Anáforas interssentenciais referem-se aos antecedentes em uma sentença diferente.

V. MÉTODOS TRADICIONAIS E CONCEITOS BÁSICOS DE RESOLUÇÃO DE ANÁFORA

Remetemos, aqui, à categorização das estratégias computacionais de resolução de anáfora feita por [7], que divide as abordagens em tradicionais ou alternativas. Os modelos tradicionais eliminam possíveis candidatos até que os mais prováveis restem. Os modelos alternativos fazem uso de técnicas da Inteligência Artificial para ranquear os candidatos mais prováveis.

O processo de determinar o antecedente de uma anáfora é chamado de resolução de anáfora.

O processo de resolução de anáfora geralmente envolve a definição de variáveis, como as características dos antecedentes e o escopo de busca. Os sistemas de resolução de anáfora, em sua maioria, têm como antecedentes os sintagmas nominais (SNs) – tarefa menos complexa em comparação a sintagmas verbais (SVs) –, e o escopo de busca é delimitado à sentença atual ou precedente.

Com base nessa definição, os SNs que precedem a anáfora são identificados como candidatos a antecedentes. A escolha do antecedente correto é feita a partir de fatores de resolução.

Os fatores tipicamente empregados são: concordância de gênero e número; restrições; consistência semântica; paralelismo sintático; paralelismo semântico; saliência; proximidade etc. Tais fatores podem ser de dois tipos: eliminação ou preferência.

Em Linguística Computacional, há divergências em relação aos termos empregados. Alguns autores defendem o emprego apenas do termo preferência.

De qualquer forma, o uso desses termos leva à distinção entre duas principais arquiteturas em resolução de anáfora: arquiteturas baseadas em restrições e arquiteturas baseadas em

preferências. A restrição é o conceito mais poderoso no processo de resolução de anáfora [8].

As restrições podem ser definidas como um conjunto de propriedades que um candidato a antecedente deve possuir. O candidato poderá ser descartado da lista de possíveis candidatos caso não possua as propriedades desejáveis.

As restrições podem estar atreladas a determinados tipos de conhecimento: sintático, semântico, pragmático etc. Um exemplo de restrição sintática é a exigência de que a anáfora e seu antecedente devem concordar em gênero e número.

A título de exemplificação, fazemos referência à teoria de ²*Centering* de Grosz e Sidner [9], um sistema de regras e restrições que define o discurso como uma estrutura hierárquica de intenções e estados de atenção constituída por segmentos. A estrutura intencional é composta por componentes globais (propósito do discurso) e componentes locais (propósito dos segmentos). O estado de atenção é composto por conjuntos de objetos, relações e propriedades no discurso.

Nos segmentos, há focos de atenção constituídos por entidades salientes. As entidades mais salientes são chamadas de centros. Os centros são objetos semânticos que podem conectar um enunciado a outros enunciados.

De acordo com [10], o conceito de centro da teoria de *Centering* é uma ferramenta importante para os fenômenos de correferência. Em resolução de anáforas interssentenciais, seu uso pode limitar grandes quantidades de candidatos. Pode ser útil também em casos ambíguos, em que o candidato mais provável pode ser o centro do segmento.

Embora a aplicação prática da teoria ainda não seja algo totalmente resolvido, tendo em vista que há vários métodos de determinar o centro de segmentos, o uso do conceito é uma constante.

A diferença entre preferências e restrições está no fato de que as preferências não são condições obrigatórias. Algumas preferências são: paralelismo sintático, paralelismo semântico e centro de atenção. O último é bastante útil em casos ambíguos, em que a decisão de escolha é baseada no candidato mais saliente, foco de atenção.

VI. PROPOSTA DE ABORDAGEM ALTERNATIVA PARA RESOLUÇÃO DE ANÁFORA

A abordagem alternativa visa à criação de dispositivos adaptativos para resolução de anáfora a partir de fatores e estratégias de métodos tradicionais e análise de *corpus* anotado.

A ideia central é a de que estudos específicos de fatores de restrição ou preferência, baseados em *corpus*, podem indicar dados interessantes, servindo para a determinação de peso, por exemplo, quando utilizados em conjunto com outros fatores, além de servirem, também, para a criação de regras para os dispositivos adaptativos e não adaptativos. É possível, ainda, que a análise dos dados sirva para o estabelecimento de fatores ainda não empregados nos métodos tradicionais.

Os principais procedimentos metodológicos a serem executados para a pesquisa são:

1. Coleta de um *corpus* na Internet de ³textos em inglês de determinados gêneros;
2. Etiquetagem automática pelo etiquetador disponível no sítio <http://beta.visl.sdu.dk/visl/en/> via programação, enviando variáveis para o formulário *online* do etiquetador e baixando as páginas de resultado;
3. Etiquetagem manual de resolução de anáforas pronominais utilizando uma ferramenta criada especificamente para a tarefa;
4. Identificação de fatores específicos para pesquisa no *corpus* processado;
5. Geração/extração automática de todos os possíveis dados interessantes em relação aos fatores identificados via programação;
6. Análise e avaliação dos dados a fim de estabelecer relações entre métodos e abordagens existentes e novos critérios de restrição ou preferência para resolução de anáforas nos textos;
7. Desenvolvimento de dispositivos adaptativos para a resolução de anáforas e montagem de exercícios;
8. Avaliação dos resultados obtidos por meio do uso de dispositivos adaptativos.

Em relação ao desenvolvimento de dispositivos adaptativos e uso da Tecnologia Adaptativa, é possível estabelecer um conjunto de regras inicial a partir de métodos existentes e análise de fatores no *corpus* anotado. Em seguida, ampliar o número de regras em uma abordagem de aprendizado automático, usando regras obtidas a partir do *corpus* anotado, sendo aprendidas em treinamento. Uma abordagem de aprendizado automático, conforme [9], evita a necessidade de realizar inúmeros testes no estabelecimento de regras.

As informações linguísticas a serem inferidas serão provenientes dos fatores obtidos pela anotação, tal como distância, concordância de gênero e número e outras a serem verificadas.

A possibilidade de uso de dispositivos adaptativos nas tarefas propostas anteriormente está baseada no trabalho de [11], que ressalta a capacidade de autômatos adaptativos serem utilizados em sistemas de aprendizado automático, especificamente no processamento de língua natural.

VII. CONSIDERAÇÕES FINAIS

A partir do estudo de fatores de restrição e preferência específicos para a resolução de anáforas com o emprego de técnicas adaptativas, esperamos apontar as possibilidades de uma abordagem alternativa que conjugue fatores de métodos tradicionais, análise de *corpus* e modelos formais. O emprego de modelos formais no desenvolvimento do SPLN para a criação automática de atividades de leitura deve minimizar inconsistências na execução de tarefas que foram observadas na investigação para o mestrado. Assim sendo, a investigação

² Centralização.

³ Textos autênticos que possam ser utilizados com o objetivo de criação de atividades de compreensão de leitura.

pode oferecer contribuições e benefícios: no âmbito da Linguística, em especial da Linguística Aplicada, para a disseminação do uso de materiais de ensino baseados em *corpora*; na interface entre a Linguística e a Computação, para o enriquecimento da interação entre a ciência da linguagem e as ciências exatas.

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.

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. Dissertação de Mestrado Inédita, LAEL, PUC-SP, 2005.
- [2] MOREIRA FILHO, P. Desenvolvimento de um software para preparação semiautomática de atividades de leitura em inglês. Dissertação de Mestrado Inédita, LAEL, PUC-SP, 2007.
- [3] DIZERÓ, W. Formalismos Adaptativos Aplicados na Modelagem de Softwares 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] KOCH, I.G.V. *A coesão textual*. 22ª ed. São Paulo: Contexto, 2010. 84 p.
- [6] Disponível em: <http://www2.jornaldacidade.net/artigos_ver.php?id=5247>. Acesso em 25/10/2011.
- [7] MITKOV, R. Anaphora resolution: the state of the art, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton, 1999.
- [8] LEFFA, V. J. A resolução de anáfora no processamento da língua natural. Relatório Final de Pesquisa da Universidade Católica de Pelotas, 2001.
- [9] NUGUES, P.M. An introduction to language processing with Perl and Prolog. Berlin: Springer-Verlag, 2006.
- [10] PARABONI, I. Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Dissertação de Mestrado Inédita. PUCRS, Porto Alegre, 1997.
- [11] MENEZES, C. E. D.; JOSÉ NETO, J. Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos. Anais da Conferência Iberoamericana em Sistemas, Cibernética e Informática, 19-21 de Julio, 2002, Orlando, Florida



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 Língua Inglesa de Oficina Pedagógica em 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