Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade

Maria das Graças V. Nunes, Sandra M. Aluisio, Thiago A. S. Pardo
NILC – ICMC – Universidade de São Paulo
São Carlos – SP, Brasil
{gracan, sandra, taspardo}@icmc.usp.br

Resumo

Este artigo faz uma breve apresentação do Núcleo Interinstitucional de Linguística Computacional (NILC), que é um dos principais grupos brasileiros dedicado a pesquisas na área de Processamento de Línguas Naturais, particularmente do português brasileiro. Após apresentar um breve histórico de sua formação, mostramos como as atuais áreas de pesquisa do grupo foram consolidando-se ao longo dos anos. Para cada uma dessas áreas de atuação do NILC, fazemos um breve resumo dos resultados mais importantes e do estado atual das pesquisas no grupo.

1. Introdução

O Núcleo Interinstitucional de Linguística Computacional (NILC)[1] é hoje composto por mais de 30 pesquisadores da área de Processamento de Línguas Naturais (PLN), incluindo professores universitários e alunos de graduação e pós-graduação, com formação principalmente em ciências da computação e linguística. Esse grupo foi criado em 1993, na Universidade de São Paulo, em São Carlos, com o objetivo de formar recursos humanos e desenvolver pesquisa e sistemas de PLN especialmente para o português do Brasil (PB). A criação do NILC foi especialmente motivada pelo convite recebido da empresa de informática Itautec, para implementar, como *plug-in* do Office da Microsoft, um sistema de revisão gramatical do português. O desafio era enorme, tendo em vista que àquela época não existiam recursos disponíveis para essa tarefa. Era necessário construir um léxico computacional, um analisador sintático robusto, voltado à detecção de erros sintáticos, e *corpora* de referência e de testes. Também era grande o desafio de compor e gerenciar uma equipe de pesquisa e desenvolvimento interdisciplinar (computação e linguística), com culturas tão distintas. Tudo isso fez com que o grupo já nascesse grande, com o compromisso de gerar um produto comercial e com a responsabilidade de criar tudo de que precisava. Apesar disso, uma primeira versão do revisor, sem análise sintática automática, foi lançada já em 1994. Outras versões se seguiram até que em 1999, por meio de uma licença que vigora até hoje, a Microsoft adquiriu direito de uso do revisor no Office 2000.

Com os recursos linguístico-computacionais construídos no projeto do revisor gramatical, até então inéditos para o PB, e já estendido com colaboradores de outras instituições – Universidade Federal de São Carlos (UFSCar) e Universidade Estadual de São Paulo (UNESP) – o grupo tornou-se referência na área de PLN e passou a ser convidado para desenvolver outros projetos, como o da Universal Networking Language (UNL). Em 1997, o NILC passou a representar o Brasil no grupo de países que integravam o Projeto UNL, patrocinado pelo Instituto de Estudos Avançados da Universidade das Nações Unidas (UNU/IAS). Mais tarde essa associação deu origem à UNDL Foundation[2], com sede em Genebra. A meta do projeto é criar ferramentas de tradução, dentro do paradigma de interlíngua, em um primeiro momento para as línguas oficiais da ONU e outras línguas de muitos falantes, para a comunicação na internet. Ao grupo brasileiro, cabia criar os recursos para a tradução entre o português e a interlíngua UNL. O projeto continua ativo na UNDL, porém, o NILC não participa mais como membro institucional. A participação, por cerca de 4 anos, no projeto UNL abriu no NILC uma importante área de pesquisa, a da tradução automática (TA). Tratava-se, à época, de uma área de pesquisa com muito pouca expressão no país. Vários outros projetos e importantes publicações têm sido gerados pelo grupo. Um relato sobre essas experiências encontra-se em (Martins *et al*., 2004a) e na Seção 4 deste artigo.

A partir do envolvimento nesses dois grandes projetos, o grupo ganhou expressão no país e no exterior e passou a agregar novos membros. Sua atuação se estendeu a áreas mais teóricas e à construção de recursos robustos para outras aplicações de PLN, o que acabou por aproximá-lo a outros grupos brasileiros de PLN.

[1] http://www.nilc.icmc.usp.br/nilc/index.html

[2] http://www.undlfoundation.org/undlfoundation/

O grupo se destaca também na organização e promoção da área de pesquisa em PLN no Brasil. Junto com outros grupos nacionais de expressão, como os da Universidade Católica do Rio Grande do Sul (PUC-RS) e do Rio de Janeiro (PUC-Rio), tem sido responsável por projetos de cooperação nacional e pelos principais eventos científicos dessa área. Esses grupos de pesquisa criaram, em 2003, o que é hoje o principal evento científico nacional dessa área, o STIL: Simpósio de Tecnologia da Informação e da Linguagem Humana[3], que está na sua 8ª edição. Da mesma forma, participam ativamente, e em conjunto com vários pesquisadores de Portugal, da organização do PROPOR[4], a conferência internacional e bianual sobre processamento do português, hoje na sua 9ª edição.

Os projetos de parceria com colegas do Brasil e do exterior têm possibilitado a geração de recursos e ferramentas de interesse de toda a comunidade e que representam avanços significativos para o processamento do PB. Podemos destacar, e detalharemos nas próximas seções, os *corpora* compilados e anotados; os diferentes léxicos computacionais; as bases e redes lexicais; ferramentas avançadas, como as que fazem análise discursiva e simplificação sintática; ferramentas aplicadas à tradução automática; aplicações como a sumarização mono e multidocumento e os ambientes de auxílio à escrita e à leitura; novos métodos de avaliação de sistemas de PLN; etc.

Atualmente o NILC conta com 14 pesquisadores seniores, de quatro diferentes instituições brasileiras, e cerca de 20 estudantes de graduação e pós-graduação associados. Sob uma perspectiva histórica, este artigo procura mostrar algumas das principais contribuições do NILC para a área de PLN no Brasil, às vésperas de completar sua maioridade, bem como apresenta um cenário das áreas atuais de atuação dos autores signatários. Na Seção 2 descrevemos brevemente os principais recursos linguístico-computacionais criados no NILC e que servem de apoio a todas as demais pesquisas. A Seção 3 apresenta os principais resultados das pesquisas do grupo na área de sistemas de auxílio à escrita e à leitura, uma das áreas de pesquisa pioneiras do NILC. A experiência do grupo em TA é relatada brevemente na Seção 4. Na Seção 5 apresenta-se a trajetória das pesquisas do grupo em sumarização automática e análise discursiva. Finalmente, na Seção 6, concluímos o artigo arriscando fazer algumas projeções para o futuro próximo.

2. Ferramentas e recursos básicos para o processamento do PT brasileiro

O primeiro recurso importante criado no NILC foi o léxico computacional (do NILC) que faz parte dos revisores ortográfico e gramatical do MS-Office. Do ponto de vista linguístico, a versão atual do léxico é capaz de gerar cerca de 1.500 mil lexemas a partir de cerca de 100 mil lemas. Cada lexema pode pertencer a uma ou mais de 13 classes, cada uma com atributos distintos. Do ponto de vista tecnológico, o léxico é implementado como um autômato finito minimizado, ocupando um espaço mínimo de memória e com desempenho otimizado (Jesus *and* Nunes, 2000). A partir do léxico, vários outros recursos lexicais foram produzidos no NILC: um tesauro eletrônico, a base Diadorim, o Unitex-BR, e finalmente a WordNet.Br.

O tesauro eletrônico TEP é resultado da primeira tentativa de se estender o léxico do NILC com informações semânticas de sinonímia e antonímia. Esse tesauro também é usado pelas ferramentas de revisão do Office para a tarefa de sugestão de alternativas. A base Diadorim é a versão do TEP disponível para a consulta na internet, na forma (ineficiente) de uma base de dados[5]. Já o Unitex-BR[6], criado segundo os formatos da ferramenta de *corpus* INTEX, é sua versão em código aberto, veiculada pela rede RELEX na web[7]. O conjunto de palavras simples no padrão DELA fez com que o número de ocorrências crescesse 93.28% em relação à fonte original. No entanto, o número de entradas do dicionário de palavras compostas, assim como o número de regras de remoção de ambiguidades, ainda é bastante tímido.

A evolução mais ambiciosa quanto à semântica lexical é a construção, em andamento, da WordNet.Br (Di Felippo *and* Dias-da-Silva, 2007; Dias-da-Silva *et al*., 2008), que segue os mesmos pressupostos da Wordnet de Princeton (Fellbaum, 1998). A versão preliminar, sob o nome TeP 2.0 (Maziero *et al*., 2008), tem interface disponível na web[8]. Atualmente, o TeP 2.0 contém 19.888 conjuntos de sinônimos e 44.678 unidades lexicais, tendo a média de 2,5 unidades por conjunto de sinônimos. Quanto à antonímia, há 4.276 relações entre os synsets da base, ou seja, aproximadamente

[3] http://www.nilc.icmc.usp.br/til/index.htm
[4] http://www.nilc.icmc.usp.br/cgpropor/
[5] http://www.nilc.icmc.usp.br/nilc/tools/intermed.htm
[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html
[7] http://infolingu.univ-mlv.fr/brasil/
[8] http://www.nilc.icmc.usp.br/tep2/index.htm

22% da base está relacionada por meio dessa relação. Além disso, para 253 unidades lexicais pertencentes à categoria dos verbos, o TeP 2.0 armazena uma frase-exemplo distinta para cada uma das unidades. A frase-exemplo fornece o contexto de uso mínimo do item lexical. O recurso armazena também uma glosa (ou seja, uma definição informal do conceito) para 6.648 synsets, todos eles constituídos por unidades da categoria dos verbos.

Toda a evolução dos recursos lexicais, bem como o desenvolvimento de ferramentas e aplicações para o processamento do português, foi acompanhada pela construção sucessiva e progressiva de diferentes *corpora*. O primeiro grande *corpus*, chamado de *corpus* NILC[9], com cerca de 40 milhões de palavras, foi compilado para subsidiar as pesquisas do revisor gramatical. Para tanto, deveria ser representativo dos desvios da língua escrita por usuários "médios" de editores de texto digital. Era preciso identificar e modelar os principais desvios gramaticais. Ao mesmo tempo, o *corpus* também deveria servir como referência para a construção de uma gramática normativa, já que a função do revisor é detectar desvios e sugerir correções. Essa dupla finalidade criou as três divisões do *corpus* NILC conhecidas como *corpus* corrigido (obras literárias, livros didáticos, textos jornalísticos, etc.), *corpus* não corrigido (redações de vestibulares) e *corpus* semi-corrigido (teses acadêmicas, cartas comerciais, etc.). O *corpus* NILC está disponível para consulta na Linguateca, no âmbito do projeto AC/DC[10].

O *corpus* NILC foi, durante muito tempo, a fonte de informação sobre o PB contemporâneo escrito para as pesquisas no grupo. A partir de 2002, com o apoio do CNPq, e em parceria com o IME (Instituto de Matemática e Estatística) e a FFLCH (Faculdade de Filosofia, Letras e Ciências Humanas), da USP-São Paulo, o projeto Lácio-Web[11], de construção de *corpora*, teve início no NILC. O objetivo deste projeto era divulgar e disponibilizar livremente na Web vários *corpora* do PB escrito contemporâneo, representando bancos de textos adequadamente compilados, catalogados e codificados em padrão de fácil intercâmbio, navegação e análise. Além disso, disponibilizar ferramentas linguístico-computacionais, tais como contadores de freqüência, etiquetadores morfossintáticos e concordanciadores. A idéia era prover recursos para um público heterogêneo: de um lado linguistas, cientistas da computação, lexicógrafos, entre outros, e, de outro, não especialistas em geral. Formado por quatro grandes *corpora*, o Lácio-Web contém 10,5 milhões de palavras de textos dos gêneros informativo, jurídico, científico, literário e instrucional.

Após o Lácio-Web, outros importantes *corpora* foram construídos pelo grupo. Destacamos a participação do NILC no Projeto do Dicionário Histórico do Português Brasileiro dos séculos 16 até o início do século 19 (HDBP)[12], tratou de várias características inerentes a textos históricos, tais como: ausência de uma ortografia, uso extensivo de abreviações e suas variações de grafia, falta de espaço entre as palavras, uso irregular da hifenização e símbolos tipográficos que caíram em desuso (Candido Jr. *et al*., 2009).

Mais recentemente, no âmbito do projeto de cooperação multi-institucional (USP, UFSCar, Unisinos, PUC-RS, PUC-Rio, Mackenzie, UNESP), o grupo coordenou a criação do Portal de *Corpus*[13] (Muniz *et al*., 2007), formado por 3 *corpora*:
(a) PLN-BR FULL, que contém 103.080 mil textos da Folha de São Paulo e 29.014.089 *tokens*; está formatado segundo etiquetas do Unitex;
(b) PLN-BR CATEG, que tem 30 mil textos e 9.780.220 *tokens*, originalmente criado para compor um *benchmark* para avaliação de métodos de classificação textual;
(c) PLN-BR GOLD, que possui 1024 textos e 338.441 *tokens* e pode ser acessado livremente via Web. O tamanho deste *corpus* é tal que representa 1% do *corpus* PLN-BR FULL de forma a conservar, proporcionalmente, a distribuição deste *corpus* maior. Trata-se de uma amostra aleatória estratificada e proporcional à distribuição do *corpus* PLN-BR FULL com relação aos textos dos cadernos do jornal. Foi criado para exemplificar e tornar pública a proposta de anotação de *corpora* da Língua Portuguesa, considerando vários níveis linguísticos (Bruckschen *et al*., 2008).

Vários outros *corpora*, de uso mais restrito a determinadas pesquisas e aplicações, têm sido compilados no NILC, como o TeMário, de sumários feitos manualmente; o *Corpus*TCC, de teses acadêmicas, e o RHETALHO, de textos acadêmicos e jornalísticos anotados pela ferramenta de análise discursiva RSTTool; o *Corpus* Paralelo, de textos alinhados português e inglês, o *corpus* paralelo de textos originais e simplificados léxica e sintaticamente, entre outros. Vários destes *corpora* são detalhados nas próximas seções.

[9] http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm
[10] http://acdc.linguateca.pt/acesso/
[11] http://www.nilc.icmc.usp.br/Lácioweb/index.htm
[12] http://www.nilc.icmc.usp.br/nilc/projects/hpc/
[13] http://www.nilc.icmc.usp.br:8180/portal/

Entre as ferramentas de PLN desenvolvidas no NILC, destacamos os etiquetadores POS construídos no âmbito do projeto Lácio-Web[14] (Aluísio *et al*., 2003), e o *parser* Curupira, que é derivado do revisor gramatical, e provê o conjunto de todas as análises sintáticas possíveis para uma dada sentença em PB (Martins *et al*. 2003).

3. Sistemas de Auxílio à Escrita e à Leitura

A necessidade de escrever artigos científicos em inglês é um dos grandes problemas de pesquisadores de vários países cuja língua nativa não é o inglês. Os trabalhos do NILC nesta área têm explorado uma estratégia de escrita baseada no reuso de trechos de textos escritos por pesquisadores nativos do inglês, indexados pelos componentes da estrutura esquemática da seção na qual aparecem.

Embora grande parte dos problemas enfrentados por escritores nativos se apresente no nível estrutural, problemas nos níveis lexical e sentencial também ocorrem. De fato, esses escritores têm o conhecimento da língua no seu uso geral, mas podem não dominar o seu uso em um gênero específico, tendo problemas na escolha de itens lexicais e estruturas sintáticas apropriadas.

Ferramentas de suporte à escrita científica em inglês com base na abordagem baseada em casos (*case-based reasoning*) e em sistemas de críticas *(expert/computer-aided critiquing systems*), largamente usados na grande área de Inteligência Artificial, foram desenvolvidas no projeto AMADEUS (AMiable Article DEvelopment for User Support) (Fontana *et al*., 1993; Aluísio *and* Oliveira, 1995; Aluísio *and* Oliveira Jr, 1996; Aluísio *and* Gantenbein, 1997; Aluísio *et al*., 2001). Estas ferramentas foram portadas para o ambiente Web, seguindo a tendência atual para facilitar o acesso de sistemas (por exemplo, o sistema SciPo-Farmácia[15] (Aluísio *et al*., 2005)), e também uma delas, chamada SciPo[16] (Feltrim, 2004; Feltrim *et al.*, 2004; Feltrim *et al.*, 2006), foi disponibilizada para a língua portuguesa para ser usada por escritores nativos do português escrevendo teses e dissertações.

Experiências realizadas com as ferramentas de auxílio à escrita científica têm demonstrado que a boa aceitação das mesmas por parte de seus usuários se deve fortemente ao fato de possuírem *corpora* específicos da área de pesquisa do usuário-escritor. Assim, uma questão que se coloca é o custo de se estender esse auxílio computacional a pesquisadores de diferentes áreas do conhecimento, pois o gargalo da construção das ferramentas é a anotação dos textos com os componentes da estrutura esquemática de um artigo, tese ou dissertação. A solução proposta no NILC foi a utilização de detecção automática dos elementos estruturais de textos científicos, dado que esta proposta se apresenta também como um desafio científico, pois trata da automatização de uma tarefa que é problemática mesmo quando realizada por humanos. Alguns sistemas têm sido propostos na literatura para a realização dessa tarefa (Burstein *et al*., 2003; Antony *and* Lashkia, 2003; Teufel *and* Moens, 2002). No NILC foram desenvolvidos dois sistemas de detecção automática de estrutura esquemática de resumos, o AZPort (Feltrim *et al.*, 2004) e o AZEA[17] (Genoves *et al*., 2007a). O primeiro é voltado para resumos em português e o segundo para resumos em inglês (*abstracts*). Ambos se baseiam no método AZ (*Argumentative Zonning*) (Teufel *and* Moens, 2002)

O SciPo (*Scientific Portuguese*), inspirado no projeto AMADEUS, é um ambiente Web voltado para escritores cuja língua mãe é o português, em especial aqueles que estão iniciando sua carreira acadêmica e ainda não estão familiarizados com as convenções do gênero científico. Ele baseia-se em teses e dissertações da área de Computação.

O SciPo apóia a estruturação e a realização linguística de textos científicos de forma flexível, deixando o usuário livre para escolher entre dois modos de trabalho, a saber: (i) um processo *top-down*, que parte do planejamento estrutural para a escrita propriamente dita, incluindo ciclos de críticas e refinamentos da estrutura, herdado do projeto AMADEUS; ou (ii) um processo *bottom-up*, em que se submete um texto já escrito à análise (detecção e crítica) automática da estrutura. Na verdade, trata-se de pontos de partida distintos para um mesmo processo cíclico de refinamento, já que a estrutura detectada e criticada em (ii) pode ser aprimorada por meio dos recursos disponíveis em (i).

O SciPo-Farmácia é um conjunto de ferramentas computacionais desenvolvido para ajudar os usuários a escreverem artigos científicos em inglês. Possui a mesma interface do SciPo, porém um

[14] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html
[15] http://www.nilc.icmc.usp.br/scipo-farmacia/
[16] http://www.nilc.icmc.usp.br/~scipo/
[17] http://www.nilc.icmc.usp.br/azea-web/

número menor de funcionalidades e baseia-se em artigos científicos da área de Ciências Farmacêuticas. Este sistema foi desenvolvido com o intuito de ajudar estudantes e pesquisadores que não têm o inglês como língua materna e necessitam escrever artigos científicos nessa língua e/ou também não estão familiarizados com a estrutura e as peculiaridades do gênero científico. O desenvolvimento do SciPo-Farmácia resultou de uma parceria entre pesquisadores da Faculdade de Ciências Farmacêuticas da USP de São Paulo e o NILC.

Outras linhas recentes de pesquisa para apoio computacional para a escrita e a leitura, no NILC, incluem:
(i) a diversificação de gênero, com o desenvolvimento de uma ferramenta Web inteligente de auxílio à escrita de planos de negócios em português (Ferraz Jr *et al.*, 2007; Raymundo *et al.*, 2007);
(ii) a implementação de uma rubrica baseada no gênero científico para analisar resumos de artigos (Aluísio *et al.*, 2005; Schuster *et al.*, 2005; Genoves *et al*, 2007a; Genoves *et al.*, 2007b);
(iii) e, mais recentemente, o desenvolvimento de tecnologias para facilitar o acesso de informação por pessoas com baixo nível de letramento ou outros problemas de leitura, no escopo do projeto PorSimples[18] (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) (Aluisio *et al*, 2008).

O grande objetivo do projeto PorSimples é poder ajudar pessoas com problemas de leitura a compreender documentos do gênero informativo disponíveis na Web brasileira, por exemplo, informações do governo e notícias de jornais de grande circulação.

No Brasil, o Indicador de Alfabetismo Funcional (INAF) tem sido computado desde 2001 para medir os níveis de letramento da população brasileira. O relatório mais atual, de 2009, apresenta um cenário ainda desanimador: 7% das pessoas são analfabetas; 21% são alfabetizadas no nível rudimentar; 47% são alfabetizadas no nível básico; e somente 25% são totalmente alfabetizadas (INAF, 2009). O número de pessoas com alfabetização nos níveis rudimentar e básico totaliza 68% da população do Brasil e estas podem somente achar informação explícita em textos curtos (rudimentares), ler e entender textos um pouco maiores, além de serem capazes de fazer inferências simples (básicas). Estes dois níveis são o alvo do projeto PorSimples, e para isso foram desenvolvidos três sistemas destinados a públicos alvos diferentes:

- um sistema de autoria, chamado SIMPLIFICA[19], para ajudar autores a produzirem textos simplificados destinados aos alfabetizados rudimentares e básicos (Candido Jr *et al.*, 2009); e
- sistemas facilitadores para ajudar o mesmo público acima a ler um dado conteúdo da Web. Estes incluem tarefas de sumarização textual e simplificação sintática (sistema FACILITA[20]) (Watanabe *et al.*, 2009) e elaboração léxica, apresentação do texto salientando as relações retóricas entre as idéias do texto, explicitação das Entidades Mencionadas e dos argumentos dos verbos (sistema FACILITA EDUCATIVO[21]) (Watanabe *et al.*, 2010).

O sistema SIMPLIFICA (Figura 1) é um editor WYSIWYG baseado no editor WEB TinyMCE[22].

Figura 1: Tela principal do SIMPLIFICA que dá acesso às 3 funcionalidades do editor: simplificação léxica e sintática (no topo, acima do texto) e verificador da inteligibilidade (na barra de status)

O usuário insere um texto no editor e realiza: (i) as escolhas para a simplificação relacionadas ao tipo de público alvo, podendo ser: simplificação forte (para alfabetizados rudimentares) em que todos os fenômenos sintáticos complexos de uma sentença são tratados; simplificação natural (para alfabetizados básicos) em que somente as sentenças apontadas por um classificador treinado em um *corpus* anotado manualmente serão tratadas; e simplificação customizada em que o usuário escolhe o fenômeno alvo de simplificação, e (ii) um ou mais tesauros a serem utilizados no processo de simplificação léxica.

[18] http://caravelas.icmc.usp.br/wiki/

[19] http://www.nilc.icmc.usp.br/porsimples/simplifica/

[20] http://vinho.intermidia.icmc.usp.br:3001/facilita/

[21] http://vinho.intermidia.icmc.usp.br/watinha/Educational-Facilita/

[22] http://tinymce.moxiecode.com/

Após as escolhas acima, o usuário pode ativar o verificador de inteligibilidade (Aluisio *et al*., 2010). Este módulo mapeia o texto em um dos 3 níveis de letramento definidos pelo INAF: rudimentar, básico, avançado. De acordo com o resultado do verificador, o usuário pode ativar simplificações léxicas e sintáticas, revisar as simplificações e iniciar novamente o ciclo, via nova checagem da inteligibilidade do texto simplificado.

O sistema FACILITA (Figura 2) é um plug-in destinado a facilitar a leitura de um documento da Web por alfabetizados dos níveis rudimentar e básico.

FACILITA inclui módulos separados de sumarização textual e simplificação sintática.

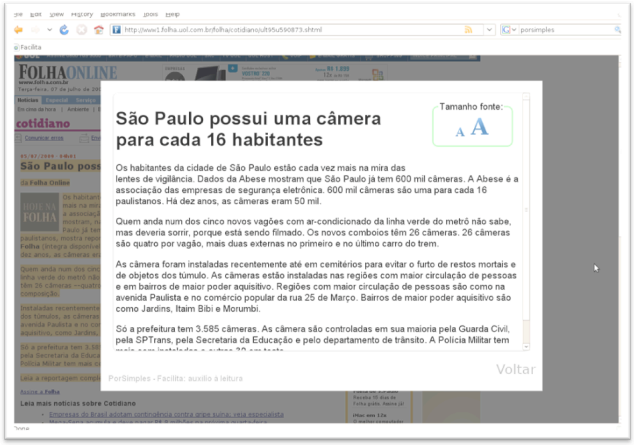

Figura 2: Janela *popup* mostrando o conteúdo facilitado de uma página Web cujo trecho em foco foi selecionado pelo usuário (ao fundo)

O usuário pode selecionar um texto de um site da Web e ativar FACILITA para obter o conteúdo facilitado. O módulo de sumarização é baseado na técnica EPC-P (extração de palavras-chaves por padrão) que verifica a presença de palavras-chaves nas sentenças do texto; aquelas que possuem palavras-chaves são retidas para o sumário final. O módulo de simplificação é melhor descrito em (Candido Jr *et al*., 2009).

O sistema Educational FACILITA[23] (Figura 3) é uma aplicação Web destinada a ajudar pessoas com baixo letramento a entenderem o conteúdo de documentos. As entidades nomeadas são marcadas e, ao serem selecionadas, definições curtas são apresentadas, vindas da Wikipédia. Também marca palavras complexas para as quais apresenta sinônimos simples.

[23] http://vinho.intermidia.icmc.usp.br/watinha/Educational-Facilita/

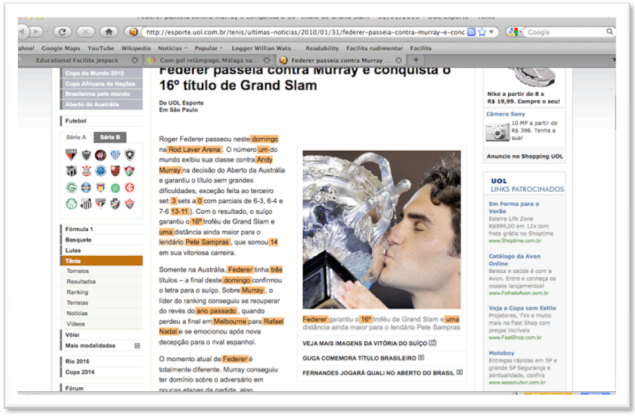

Figura 3: Resultado do sistema de elaboração textual FACILITA EDUCACIONAL ao ser acionado de uma página Web.

Mais detalhes dos recursos, métodos, sistemas e ferramentas de suporte disponibilizados pelo PorSimples podem ser vistos em Aluisio *and* Gasperin (2010).

4. Tradução Automática

O NILC possui trabalhos em TA de diferentes paradigmas. Com o projeto UNL, tiveram origem alguns trabalhos em TA por interlíngua (Martins *et al*., 2004a). O projeto Retratos, por sua vez, investiga a tradução entre português, espanhol e inglês, por meio de regras de tradução aprendidas automaticamente de *corpus*. Já no paradigma estatístico, alguns trabalhos procuram criar os primeiros recursos para PB. Outros trabalhos relacionados a TA dizem respeito à desambiguação lexical e à avaliação de traduções automáticas. Comentamos a seguir sobre essas linhas de pesquisa.

O projeto EPT-Web[24] propôs um sistema de tradução por interlíngua de *headlines* de notícias do The NewYork Times para o português. Para a criação do protótipo foi necessário criar um conjunto interessante de recursos: um dicionário trilíngue inglês-UNL-português (Antiqueira *et al*., 2002), um sistema de tradução inglês-UNL (Martins *et al*., 2004b) e um sistema de tradução UNL-português, este derivado dos trabalhos desenvolvidos no Projeto UNL.

Outra experiência com a UNL ocorreu no projeto LIBRAS[25], que visava a tradução de PB para a língua brasileira de sinais, Libras. Resultados preliminares evidenciaram a complexidade de se

[24] http://www.nilc.icmc.usp.br/nilc/projects/ept-web.htm
[25] http://www.nilc.icmc.usp.br/nilc/projects/LIBRAS2.htm

relacionar semanticamente 3 línguas de naturezas distintas: uma língua natural (PB), uma língua gestual-visual (Libras), e uma interlíngua (UNL) cuja função é representar a semântica comum entre as outras duas (Nunes *et al*., 2003).

Na linha da tradução baseada em regras, o projeto Retratos (Caseli, 2007) desenvolveu ferramentas de alinhamento sentencial e lexical para as línguas portuguesa, inglesa e espanhola, criou *corpora* paralelos e sistemas de aprendizagem automática de léxicos bilíngues e de regras de tradução (Nunes *et al*., 2008). Este projeto compartilhou alguns recursos do sistema de tradução de código aberto, Apertium[26]. Os recursos criados têm servido para apoiar outras pesquisas, como a de reconhecimento de multipalavras a partir de *corpus* paralelo bilíngue (Caseli *et al*., 2009).

Pelo que se sabe, os primeiros trabalhos no Brasil na linha da TA estatística foram realizados no NILC. Aziz *et al*. (2008) desenvolveram um tradutor estatístico entre o PB e o espanhol. Com base em um *corpus* paralelo relativamente pequeno de notícias de divulgação científica da Revista Pesquisa FAPESP, treinaram-se alguns modelos estatísticos clássicos baseados em palavras (Brown *et al*., 1993). Os resultados obtidos foram levemente inferiores ao tradutor Apertium. Dando continuidade a este trabalho, Aziz *et al*. (2009a) treinaram modelos estatísticos mais sofisticados baseados em *phrases* (que, nesse contexto, significam sequências quaisquer de palavras) (Koehn *et al*., 2003), incluindo, além das línguas anteriores, o inglês americano. Utilizando-se o mesmo *corpus*, os resultados obtidos foram superiores aos obtidos pelo Apertium para o par de línguas português-espanhol. Por meio de um experimento preliminar, constatou-se que os resultados são comparáveis ao Google Translate[27] para o par de línguas português-inglês.

Além dos trabalhos anteriores, trabalhos complementares de Caseli e Nunes (2009), Nunes e Caseli (2009) e Aziz *et al*. (2009b) investigaram como alguns parâmetros e simples escolhas de modelagem podem interferir na qualidade da tradução produzida pelos métodos de TA estatística. Por exemplo, investigaram-se as questões de uniformização de fonte, uso de pontuação no texto, e aplicação de otimização dos valores de parâmetros estatísticos, dentre outros, demonstrando-se que algumas pequenas alterações podem influenciar positivamente os resultados.

[26] http://www.apertium.org/
[27] http://translate.google.com/

O projeto LeAR investigou a desambiguação lexical de sentido (WSD) para a TA. Propôs uma nova abordagem de WSD voltada especificamente para a tradução automática, que segue uma metodologia híbrida - baseada em conhecimento e em *corpus* - e utiliza um formalismo relacional para a representação de vários tipos de conhecimento e de exemplos de desambiguação, por meio da técnica de Programação Lógica Indutiva (ILP). Experimentos diversos mostraram que a abordagem proposta supera abordagens alternativas para a desambiguação multilíngue e apresenta desempenho superior ou comparável ao do estado da arte em desambiguação monolíngüe. Adicionalmente, tal abordagem se mostrou efetiva como mecanismo auxiliar para a escolha lexical na tradução automática estatística (Specia *et al*., 2009a). Este trabalho também mostrou como a ILP, juntamente com vários tipos de conhecimento de fundo, podem melhorar consideravelmente o desempenho de sistemas de desambiguação lexical de sentido (Specia *et al*., 2009b)

Outra linha relacionada à TA é a que investiga métodos alternativos para avaliação automática de traduções automáticas. O estabelecimento de métricas para avaliação automática da qualidade dos sistemas de tradução automática é crucial devido ao amplo uso da TA na web, e isto pode ser feito representando-se textos como redes complexas. Os conceitos e metodologias de redes complexas vêm sendo usados numa enorme variedade de áreas (Costa *et al*., 2008), incluindo a análise automática de textos em PLN. O potencial uso de redes complexas para esse tipo de análise foi demonstrado em várias oportunidades, a partir da comprovação de que um texto pode ser representado por uma rede livre de escala (Cancho *and* Sole, 2001), isto é, uma rede com poucos vértices fortemente conectados e muitos vértices fracamente conectados. Resultados consolidados no grupo incluem a determinação de autoria (Antiqueira et. al., 2007), a avaliação da qualidade de sumários automáticos (Antiqueira *et al*., 2009), e de tradução automática (Amancio *et al*., 2008). Neste último cenário, métricas de redes complexas foram aplicadas e os resultados foram utilizados como entrada para métodos de aprendizado de máquina, e permitiram que textos traduzidos automaticamente e manualmente fossem distinguidos. Tal método foi aplicado para o par de línguas inglês-português e espanhol-português. Os resultados mostram que é possível capturar um contexto mais amplo com a utilização de níveis hierárquicos mais profundos em conjunto com os métodos de aprendizado de máquina.

5. Sumarização mono e multidocumento

Há tradição no NILC em trabalhos de sumarização automática, principalmente em sumarização monodocumento. Há trabalhos tanto da abordagem profunda baseados em teorias discursivas como baseados em aprendizado de máquina e métodos empíricos. Pelo que se sabe, o NILC é o único grupo de pesquisa no Brasil que desenvolve pesquisas nesse assunto.

O primeiro trabalho de sumarização foi teórico e com base em conhecimento discursivo, realizado por Rino (1996) e validado na forma de um gerador automático de sumários por Pardo e Rino (2002). Estes trabalhos foram baseados na combinação de 3 modelos discursivos: a Teoria de Estruturação Retórica RST (Mann e Thompson, 1987), o modelo intencional de Grosz e Sidner (1986) e o modelo Problema-Solução (Jordan, 1980). Combinando-se o conhecimento fornecido por esses três modelos, gerava-se o sumário de textos científicos. Essa abordagem produziu bons resultados, apesar de ser altamente custosa devido à demanda por conhecimento muito especializado.

Outros trabalhos baseados somente em RST seguiram os trabalhos anteriores. O melhor representante desta linha talvez seja o trabalho de Uzêda *et al*. (2008), onde se analisaram diversos métodos de sumarização com base na RST e se demonstrou que todos eles têm desempenho comparável. Mostrou-se também que os métodos baseados em RST são melhores do que métodos superficiais clássicos.

Ainda nesta linha, Seno e Rino (2005) e Carbonel *et al*. (2007) investigaram o uso da Teoria das Veias (Cristea *et al*., 1998) para lidar com correferências em sumários, já que a ocorrência de anáforas não resolvidas em sumários provoca sérios problemas de coesão e coerência. A Teoria das Veias é um modelo que permite que se identifiquem os segmentos textuais possíveis em que antecedentes de anáforas ocorram, o que possibilitaria a inclusão destes segmentos no sumário, resolvendo a anáfora e melhorando sua qualidade, portanto. Esse modelo, entretanto, trabalha sobre estruturas RST, demandando novamente conhecimento especializado. Como a Teoria das Veias indica várias possibilidades para a ocorrência do antecedente de uma anáfora, Tomazela e Rino (2009) investigaram como informação semântica superficial (de nível lexical) pode ajudar neste processo. Sua hipótese principal foi que o antecedente deve apresentar os mesmos traços semânticos da anáfora, o que permitira descartar algumas possibilidades de segmentos fornecidas pela Teoria das Veias.

Em outra linha, mas ainda na abordagem profunda, Martins e Rino (2002) usaram uma interlíngua para representar o conteúdo textual e manualmente desenvolveram regras para produzir suas versões comprimidas. A interlíngua utilizada foi a UNL, já citada na seção anterior.

É importante notar que muitos dos sistemas citados anteriormente se baseiam na RST. Diante desta demanda, foi produzido para o PB um analisador automático chamado DiZer (Pardo e Nunes, 2008). Esse analisador, de natureza simbólica (com regras de análise produzidas manualmente a partir de estudo de *corpus*) produz as estruturas RST possíveis para um texto-fonte de entrada. Como esse analisador foi desenvolvido para textos científicos em português e era de difícil adaptação para outros tipos textuais e línguas, novos trabalhos foram iniciados e se desenvolveu o DiZer 2.0[28], que está *online* e consiste em uma solução web de mais fácil portabilidade para outras línguas e tipos textuais. Esta versão do analisador permite que um usuário de forma relativamente simples adicione os recursos necessários e personalize seu próprio analisador.

Na abordagem superficial, Pardo *et al*. (2003a) desenvolveram um sistema de sumarização baseado principalmente em freqüência de palavras. Esse sistema é provavelmente um dos sistemas superficiais mais usados no Brasil e, apesar dos sumários gerados apresentarem diversos problemas de coesão e coerência, seus resultados são interessantes. Trabalhando sobre esses resultados, Gonçalves *et al*. (2008) usaram regras de pós-processamento para resolver anáforas e demonstraram que muitos dos problemas anteriores eram resolvidos.

Pardo *et al*. (2003b) usaram uma rede neural de Kohonen e atributos superficiais para modelar o processo de sumarização. O princípio deste trabalho consistia em agrupar sentenças de igual importância por meio da rede treinada, de forma que fosse possível descartar sentenças menos importantes para a produção do sumário.

Leite *et al*. (2007, 2008) usaram, para seu sumarizador, um método de aprendizado de máquina bayesiano para combinar atributos superficiais simples e complexos, produzindo os melhores resultados até o momento para a língua portuguesa. Um dos pontos interessantes deste trabalho é que seus atributos complexos codificam

[28] http://www.nilc.icmc.usp.br/dizer2

métodos completos de sumarização automática, atribuindo, desta forma, grande informatividade ao processo como um todo.

Antiqueira *et al*. (2009) modelaram textos como redes complexas e usaram métricas das redes para selecionar informação relevante para compor o sumário, produzindo resultados muito bons. Sua modelagem de texto como rede é muito simples e elegante, demonstrando que não é necessária grande quantidade de conhecimento linguístico para se gerar bons sumários.

Trabalhos mais antigos do grupo incluem as propostas de Souza e Nunes (2001) e Pereira *et al*. (2002), as quais também usaram atributos textuais superficiais para sumarização. Outra questão relacionada investigada foi a compressão sentencial, ou seja, a tarefa de se produzir uma versão mais curta de uma sentença (Kawamoto e Pardo, 2010), utilizando-se aprendizado de máquina. Tal abordagem investigou o aprendizado automático de regras simbólicas para detecção de palavras de uma sentença que poderiam ser excluídas, observando-se critérios de gramaticalidade, informatividade e foco textual.

Recentemente, sistemas de sumarização multidocumento começaram a ser investigados no Brasil. O primeiro sistema foi proposto por Pardo (2005) e era trivial: o sistema simplesmente justapõe todos os textos e aplica métodos de seleção de sentenças com base na freqüência das palavras. Desde 2009, um grande projeto de sumarização multidocumento da abordagem profunda foi iniciado. Com base no modelo CST (*Cross-document Structure Theory*) (Radev, 2000), diversas estratégias de sumarização estão sendo investigadas, com alguns resultados promissores já produzidos. A CST, inspirada na RST, modela o relacionamento entre diversos textos sobre um mesmo assunto, permitindo que se lide adequadamente com os fenômenos multidocumento, como a presença de informação redundante, contraditória e complementar, a ordenação das informações textuais no sumário, e a própria questão de coerência e coesão.

Os primeiros trabalhos nesta linha de sumarização multidocumento (Jorge e Pardo, 2009, 2010) relacionaram preferências de sumarização do usuário com os relacionamentos previstos na CST, produzindo operadores de sumarização que, quando aplicados ao conteúdo textual, produzem um ranque de informações a partir do qual se devem selecionar as que serão incluídas no sumário.

Novamente, devido à demanda por análise CST, investiga-se atualmente a questão da análise automática multidocumento segundo este modelo. Os primeiros resultados obtidos (usando aprendizado de máquina e atributos superficiais) são promissores e avançam significativamente o estado da arte (Maziero *et al*., 2010).

Durante a investigação da sumarização automática no NILC, diversos recursos e ferramentas dedicados ao assunto foram produzidos. Dentre os *corpora*, os de mais destaque são o TeMário (Pardo e Rino, 2003; Maziero *et al*., 2007), o CSTNews (Aleixo e Pardo, 2008), o Summ-it (Collovini *et al*., 2007) e o Rhetalho (Pardo e Seno, 2005). Em termos de ferramentas, valem citar a RST Toolkit e a CSTTool, que são ferramentas de suporte à análise RST e CST, respectivamente.

6. Conclusões

Nos últimos 17 anos, o NILC tem se dedicado à pesquisa e ao desenvolvimento de recursos e sistemas de PLN, especialmente para o PB escrito. Ao contrário do cenário inicial, hoje já é possível desenvolver pesquisa em qualquer área de PLN para o português em condições competitivas com outras línguas. Recursos básicos como léxicos, *corpora*, *parsers* e modelos de língua estão ao alcance dos pesquisadores, e o NILC se orgulha de ter contribuído significativamente para isto. Os desafios, no entanto, continuam grandes. É necessário fazer crescer a comunidade de PLN no país, que atualmente encontra dificuldades decorrentes do modelo de educação superior formal. Um linguista encontra barreiras para complementar sua formação em Computação, da mesma forma que um cientista da computação as encontra para complementar a sua em Linguística. Essa formação híbrida tem acontecido de maneira quase *ad hoc*, o que impede uma formação continuada. Para alterar o modelo, no entanto, é preciso fortalecer a área, inicialmente dentro dos limites de ambas as comunidades, e posteriormente além deles. Esse fortalecimento decorre de pesquisas de boa qualidade e reconhecidas internacionalmente, bem como de uma comunidade local unida e com objetivos comuns. Nesse sentido, ações como a organização dessa comunidade em comissões especiais (como a Comissão Especial de PLN na Sociedade Brasileira de Computação[29]), e a promoção de eventos científicos para atrair novos pesquisadores (como a Escola Brasileira de Linguística Computacional[30]), a aproximação a sociedades internacionais, como a ACL e a NAACL, são muito relevantes.

[29] http://www.nilc.icmc.usp.br/cepln/
[30] http://www.corpuslg.org/ebralc/Inicial.html

Do ponto de vista das pesquisas do NILC, o momento atual é de consolidação de trabalhos iniciados há bastante tempo, como os de Ferramentas de Auxílio à Escrita e à Leitura e os de Sumarização Automática.

Na linha de trabalhos sobre ferramentas de suporte à escrita científica, o foco de pesquisa para os próximos anos será estender a rubrica baseada em gênero científico para uso em outras seções além do resumo. Quando totalmente automatizada, esta rubrica possibilitará que uma ferramenta de suporte à escrita detecte erros e ofereça sugestões para melhorias.

Quanto aos trabalhos dentro do escopo do projeto PorSimples, trabalhos futuros focarão na avaliação das ferramentas com usuários reais. Também pretende-se melhorar o desempenho da simplificação sintática via experimentos com *parsers* sintáticos de abordagens diferentes do atual utilizado no projeto.

Sobre os trabalhos de sumarização automática, é interessante notar sua evolução natural. No início, o grupo investia pesadamente em abordagens profundas, necessitando de ferramentas de análise sofisticadas. Atualmente, tais ferramentas já existem (mesmo que ainda longe de produzirem dados ideais) e a transição entre as investigações monodocumento para multidocumento foi iniciada. Na linha superficial, resultados do estado da arte foram atingidos, incentivando a continuidade das investigações nesta direção.

Os trabalhos em tradução automática têm se concentrado cada vez mais na linha estatística, mas não abandonando o uso de conhecimento linguístico. Investigações recentes procuram saber como o conhecimento sintático-semântico pode auxiliar nesse processo. Acredita-se que, como na maior parte das aplicações de PLN, a combinação das abordagens pode produzir resultados melhores.

Agradecimentos

Agradecemos a todos os colaboradores do NILC, desde sua criação, que têm tornado possível o desenvolvimento de todos os trabalhos - entre muitos outros - descritos neste artigo. Agradecemos também o apoio das agências brasileiras de pesquisa – CNPq, FAPESP, CAPES e FINEP –, da UNU/IAS e da Itautec S.A.

Referências

Aleixo, P. e Pardo, T.A.S. 2008. *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory).* Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, Maio, 12p.

Aluísio, S. M.; Pelizzoni, J. M.; Marchi, A. R.; Oliveira, L. H.; Manenti, R.; Marquivafável, V. 2003. An account of the challenge of tagging a reference *corpus* of Brazilian Portuguese In: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003, Proceedings. *Lecture Notes in Computer Science 2721* Springer 2003

Aluísio, S. M., Barcelos, I., Sampaio, J., Oliveira Jr, O. N. 2001. How to Learn the Many Unwritten ´Rules of the Game´ of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing In: *IEEE International Conference on Advanced Learning Technologies,* Madison, Wisconsin. 2001. v.1. p.257 – 260.

Aluísio, S. M., Fontana, N., Oliveira JR., O. N., Oliveira, M. C. F. 1993. Computer Assisted Writing - Applications to English as a Foreign Language. *Computer Assisted Language Learning Journal. v*.6, p.145 - 161, 1993.

Aluísio, S. M., Gantenbein, R. E. 1997. Towards the Application of Systemic Functional Linguistics in Writing Tools In*: Proceedings of International Conference on Computers and their Applications,* 1997. v.1. p.181 - 185

Aluísio, S. M., Oliveira JR, O. N. 1995. A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users In: *Proceedings of the First International Conference - ICCBR-95. Lecture Notes in Artificial Intelligence.* Berlin: Springer-Verlag, v.1010. p. 121 – 132

Aluísio, S. M., Oliveira JR., O. N. 1996. Detailed Schematic Structure of Research Papers Introductions: An Application in Support-Writing Tools. *Revista de La Sociedad Espanyola Para El Procesamiento Del Lenguage Natura.* v.1, p.141 – 147.

Aluísio; S. M.; Schuster; E.; Feltrim; V.D.; Pessoa Jr; A.; Oliveira JR, O. N. 2005. Evaluating scientific abstracts with a genre-specific rubric. In: *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005).* Amsterdam: v.1, p. 738-740.

Aluísio, S. M., Specia, L., Pardo, T.A.S., Maziero, E. G. *and* Fortes, R. 2008. Towards Brazilian

Portuguese Automatic Text Simplification Systems. In the *Proceedings of the 8th ACM Symposium on Document Engineering*, pp. 240-248.

Aluisio, S., Specia, L., Gasperin, C. *and* Scarton, C. 2010. Readability Assessment for Text Simplification. To be published in the Proceedings of the The *5th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2010.*

Aluísio, S.M. *and* Gasperin, C. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. To be published in the *Proceedings of The Young Investigators in the Americas Workshop, NAACL-HLT 2010.*

Amancio, D.R.; Antiqueira, L.; Pardo, T.A.S.; Costa, L.F.; Oliveira Jr. O.N.; Nunes, M.G.V. 2008. Complex networks analysis of manual *and* machine translations. *International Journal of Modern Physics C - IJMPC*, V. 19, N. 4, pp. 583-598.

Anthony, L., Lashkia, G.V. 2003. Mover: A machine learning tool to assist in the reading *and* writing of technical papers. *IEEE Transactions on Professional Communication* 46 (2003) 185-193

Antiqueira, L ; Oliveira Jr, Osvaldo N.; Costa, Luciano F.;; Nunes, M. G. V. 2009. A complex network approach to text summarization. *Information Sciences,* v. 179, p. 584-599.

Antiqueira, L.; Pardo, T. A. S.; Nunes, M. G. V.; Oliveira Jr., O. N. 2007. Some issues on complex networks for author characterization. *Inteligencia Artificial*, v. 11, p. 51-58.

Antiqueira L.; Fossey, M.F.; Pedrolongo, T.; Greghi, J.G.; Martins, R.T.; Nunes, M.G.V. 2002. *A construção do corpus e dos dicionários Inglês-UNL e UNL-português para o projeto EPT-Web* - Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - NILC-TR-02-24.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2009a. Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese. In *Anais do VII Encontro Nacional de Inteligência Artificial* - ENIA, pp. 769-778. July 20-24, Bento Gonçalves/RS, Brazil.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2009b. Fine-tuning in Portuguese-English Statistical Machine Translation. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-4. September 8-10, São Carlos/SP, Brazil.

Aziz, W.F.; Pardo, T.A.S.; Paraboni, I. 2008. An Experiment in Spanish-Portuguese Statistical Machine Translation. In the Proceedings of the 19th Brazilian Symposium on Artificial Intelligence - SBIA *(Lecture Notes in Computer Science 5249*), pp. 248-257. Salvador-BA, Brazil. October, 26-30.

Brown, P.E.; Pietra, S.A.D.; Pietra, V.J.D.; Mercer, R.L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 16, N. 2, pp. 79-85.

Bruckschen, M.; Muniz, F.; Souza, J. G. C.; Fuchs, J. T.; Infante, K.; Muniz, M.; Gonçalves, P. N.; Vieira, R.; Aluísio, S. M. 2008. *Anotação Linguística em XML do Corpus PLN-BR*. Série de Relatórios do NILC (NILC-TR-09-08). São Carlos - SP, Junho 2008, 39 p.

Burstein, J.; Marcu, D.; Knight, K. 2003. Finding the WRITE Stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1):32–39.

Candido Jr., A., Aluísio, S. M. 2009. Building a *Corpus*-based Historical Portuguese Dictionary: Challenges *and* Opportunities. *Traitement Automatique des Langues (TAL),* [S.l.], v.50, p.73 – 102. ISSN: 1965-0906

Carbonel, T.I.; Pelizzoni, J.; Rino, L.H.M. 2007. Validação Preliminar da Teoria das Veias para o Português e Lições Aprendidas. In the *Proceedings of the V Workshop on Information and Human Language Technology.* Rio de Janeiro-RJ.

Caseli, H.M. *and* Nunes, I.A. 2009. Statistical Machine Translation: little changes big impacts. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL. September 8-10, São Carlos/SP, Brazil.

Caseli, H.M.; Ramisch C.; Nunes, M.G.V.; Villavicencio, A. 2009. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, p. 1-19.

Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. 2008. Automatic induction of bilingual resources from aligned parallel *corpora*: application to shallow-transfer machine translation. *Machine Translation.* v. 1, p. 227-245.

Caseli, H.M. 2007. *Indução de léxicos bilíngues e regras para a tradução automática*. Tese de Doutorado. ICMC-USP, Abril, 2007. 158 p.

Collovini, S.; Carbonel, T.I.; Fuchs, J.T.; Coelho, J.C.B.; Rino, L.H.M.; Vieira, R. 2007. Summ-it: Um *corpus* anotado com informações discursivas visando à sumarização automática. In the

Proceedings of the V Workshop on Information and Human Language Technology. Rio de Janeiro/RJ.

Costa, L. F.; Oliveira Jr., O. N.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P. R.; Antiqueira, L.; Viana, M. P.; Rocha, L. E. C. 2008. Analyzing *and* Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *Physics and Society*.

Cristea, D.; Ide, N.; Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion *and* Coherence. In the *Proceedings of the Coling-ACL*, pp. 281-285. Montreal, Canadá.

Di Felippo, A. and Dias-da-Silva, B.C. 2007. Towards na automatic strategy for acquiring the WordNet.Br hierarchical relations. In *Proceedings of the 5th Workshop in Information and Human Language Technology.* Rio de Janeiro, Brasil.

Dias-da-Silva, B.C.; Di Felippo, A. and Nunes, M.G.V. 2008. The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

Fellbaum, C. 1998. *WordNet: an electronic lexical database*. Ca., MA: MIT Press.

Feltrim, V., Aluisio, S.M., Nunes, M.G.V.2003. Analysis of the rhetorical structure of computer science abstracts in portuguese. In Archer, D., Rayson, P., Wilson, A., McEnery, T., eds.: *Proceedings of Corpus Linguistics 2003*, *UCREL Technical Papers,* Vol. 16, Part 1, Special Issue. (2003) 212-218

Feltrim, V. D. 2004. *Uma Abordagem baseada em Corpus e em Sistemas de Crítica para a onstrução de Ambientes Web de Auxílio à Escrita Acadêmica em Português*. Tese de Doutorado. ICMC – USP, São Carlos, 181p.

Feltrim, V. D., Pelizzoni, J. M., Teufel, S., Nunes, M. G. V., Aluisio, S.M. 2004. Applying Argumentative Zoning in an automatic critiquer of academic writing. In Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004), *Lecture Notes in Artificial Intelligence*, 3171, Springer, p. 214-223.

Feltrim, V., Teufel, S., Nunes, M.G.V., Aluísio, S. M. 2006. Argumentative Zoning Applied to Critiquing Novices'ScientificAbstracts In: *Computing Attitude and Affect in Text: Theory and Applications*. Ed. Dordrecht, The Netherlands : Springer, 2006 v.1, p. 159-170.

Ferraz Jr, C.C.P., Boas, E.V.B., Dornelas, J., Amancio, M.A., Raymundo, E., Aluísio, S. M., Feltrim, Valéria D. 2007. PlaNInt!: Uma ferramenta Web inteligente de auxílio à escrita de planos de negócios em português. *Locus Científico.* v.1, p.48 - 57.

Fontana, N.; Aluísio, S.M.; Oliveira, M.C.F.; Oliveira JR., O.N. 1993. Computer assisted writing - applications to English as a foreign language. 145-161. *CALL (Computer Assisted Language Learning Journal),* 6, 145-161.

Genoves JR, Luiz Carlos, Feltrim, Valéria D., Dayrell, C., Aluísio, S. M. 2007a. Automatically detecting schematic structure components of English abstracts: building a high accuracy classifier for the task. In: *International Workshop on Natural Language Processing for Educational Resources in conjunction with the International Conference RANLP´2007*, 2007, Borovets, v.1. p.23 – 29.

Genoves JR, L.C., Lizotte, R., Schuster, E., Dayrell, C., Aluísio, S. M. 2007b. A two-tiered approach to detecting English article usage: an application in scientific paper writing tools In: *Proceedings of the RANLP-2007,* Sofia: Bulgarian Academy of Sciences, 2007. v.1. p.225 – 229.

Gonçalves, P.N.; Vieira, R.; Rino, L.H.M. 2008. CorrefSum: Referencial Cohesion Recovery in Extractive Summaries. *Lecture Notes in Artificial Intelligence* (Proc. of the 8th International Conference on Computational Processing of Portuguese Language, Propor2008). Berlin : Springer, 2008. v. 5190. p. 224-227.

Grosz, B. *and* Sidner, C. 1986. Attention, Intentions, *and* the Structure of Discourse. *Computational Linguistics*, Vol. 12, No. 3.

INAF 2009. Instituto P. Montenegro e Ação Educativa. INAF Brasil - Indicador de Alfabetismo Funcional - 2009. Disponível em: http://ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf

Jesus, M.A.C.; Nunes, M.G.V. 2000. Autômatos Finitos e Representação de Grandes Léxicos: Aplicação a um Léxico de Português Brasileiro. In *Anais do V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000),* v.1, p.29-42.

Jordan, M.P. 1980. Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252.

Jorge, M.L.C. and Pardo, T.A.S. 2009. Content Selection Operators for Multidocument Summarization based on Cross-document Structure

Theory. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-8. September 8-10, São Carlos/SP, Brazil.

Jorge, M.L.C. *and* Pardo, T.A.S. 2010. Formalizing CST-based Content Selection Operations. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language* - PROPOR. April, 27-30, Porto Alegre/RS, Brazil.

Junior, A. C., Maziero, E., Gasperin, C., Pardo, T., Specia, L.*;* Aluisio, S. M. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In the *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.

Kawamoto, D. and Pardo, T.A.S. 2010. Learning Sentence Reduction Rules for Brazilian Portuguese. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* – NLPCS. Funchal, Madeira, Portugal.

Koehn, P.; Och, F.J.; Marcu, D. 2003. Statistical phrase-based translation. In the *Proceedings of the HLT-NAACL*, pp. 48-54.

Leite, D.S.; Rino, L.H.M.; Pardo, T.A.S.; Nunes, M.G.V. 2007. Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In C. Biemann, I. Matveeva, R. Mihalcea, *and* D. Radev (eds.), *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pp.17-24. Rochester, NY, USA.

Leite, D.S. and Rino, L.H.M. 2008. Combining Multiple Features for Automatic Text Summarization through Machine Learning. In *Lecture Notes in Artificial Intelligence* (Proc. of the 8th International Conference on Computational Processing of Portuguese Language, Propor2008), 2008. v. 5190. p. 122-132.

Mann, W.C. *and* Thompson, S.A. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Martins, C.B. and Rino, L.H.M. 2002. Revisiting UNLSumm: Improvement through a case study. In the *Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing*, Vol. 1. pp. 71-79. Sevilha, Espanha.

Martins, R.T.; Pelizzoni, J.M.; Hasegawa, R; Nunes, M. G. V. 2004a. Da tradução automática para a língua portuguesa: apontamentos de três experiências baseadas em interlíngua. *Palavra (PUCRJ),* Rio de Janeiro, v. 12, n. 1, p. 37-55.

Martins, R.T., Hasegawa, R., Nunes, M. G. V. 2004b. HERMETO: A Natural Language Analysis Environment In: TIL- Workshop em Tecnologia da Informação e da Linguagem Humana, 2004, Salvador. *Anais do SBC 2004*.

Martins, R. T.; Hasegawa, R.; Nunes, M.G.V. 2003. Curupira: a functional *parser* for Brazilian Portuguese. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes (Eds.): Proceedings of the Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. *Lecture Notes in Computer Science 2721* Springer 2003, ISBN 3-540-40436-8.

Maziero, E.G.; Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. 2007. *TeMário 2006: Estendendo o Córpus TeMário*. Série de Relatórios do NILC. NILC-TR-07-06. São Carlos-SP, Agosto, 8p.

Maziero E.G.; Jorge, M.L.C.; Pardo, T.A.S. 2010. Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* – NLPCS. Funchal, Madeira, Portugal.

Maziero, E.G., Pardo, T.A.S., Di Felippo, A., Dias-da-Silva, B.C. 2008. A Base de Dados Lexical e a Interface Web do TeP 2,0 - Thesaurus Eletrônico para o Português do Brasil. *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp, 390-392.

Muniz, M.; Paulovich, F. V.; Minghim, R.; Infante, K.; Muniz, F.; Vieira, R.; Aluísio, S. 2007. Taming the tiger topic: an XCES compliant *corpus* Portal to generate sub*corpus* based on automatic text topic identification. In: *Proceedings of the Corpus Linguistics 2007 Conference*.

Nunes, M.G.V., Pelizzoni, J. M., Greghi, J. G., Hasegawa, R., Martins, R. T. 2003. *Projeto PULO*. NILC Project Report, Jun. 2003

Nunes, I.A. e Caseli, H.M. 2009. Primeiros Experimentos na Investigação e Avaliação da Tradução Automática Estatística Inglês-Português. Em *Anais do Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana* – TILic. São Carlos, Brasil.

Pardo, T.A.S. and Rino, L.H.M. 2002. DMSumm: Review *and* Assessment. In E. Ranchhod and N. J. Mamede (eds.), 3rd International Conference: Portugal for Natural Language Processing –

PorTAL (*Lecture Notes in Artificial Intelligence 2389*), pp. 263-273. Faro, Portugal. June 23-26.

Pardo, T.A.S. e Rino, L.H.M. 2003. *TeMário: Um Corpus para Sumarização Automática de Textos*. Série de Relatórios do NILC. NILC-TR-03-09. São Carlos-SP, Outubro, 13p.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. 2003a. GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*. Faro, Portugal.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. 2003b. NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. In *Anais do IV Encontro Nacional de Inteligência Artificial* – ENIA, pp. 1-10. Campinas-SP, Brazil.

Pardo, T.A.S. 2005. *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, Fevereiro, 8p.

Pardo, T.A.S. e Seno, E.R.M. 2005. Rhetalho: um *corpus* de referência anotado retoricamente. In *Anais do V Encontro de Corpora*. São Carlos-SP, Brasil. 25 a 26 de Novembro.

Pardo, T.A.S. *and* Nunes, M.G.V. 2008. On the Development *and* Evaluation of a Brazilian Portuguese Discourse *Parser*. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43-64.

Pereira, M.B.; Souza, C.F.R.; Nunes, M.G.V. 2002. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. *Revista Eletrônica de Iniciação Científica*. Ano II, Vol. 2, N. 1.

Radev, D.R. 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Raymundo, E., Amancio, M.A., Feltrim, Valéria D., Aluísio, S. M. 2007. Análise da Estrutura Retórica da Seção Sumário Executivo de Plano de Negócios In: *Anais do VI Encontro de Linguística de Corpus*, p.1 – 18.

Rino, L.H.M. 1996. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP. São Carlos - SP.

Seno, E.R.M. *and* Rino, L.H.M. 2005. Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In the *Proceedings of the RANLP 2005 Workshop on Crossing Barriers in Text Summarization Research*, pp. 70-75.

Souza, C.F.R. and Nunes, M.G.V. 2001. *Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português*. Technnical Report NILC-TR-01-09.

Specia, L.; Nunes, M.G.V.; Stevenson, M. 2009a. Assessing the contribution of shallow *and* deep knowledge sources for word sense disambiguation. *Language Resources and Evaluation*, Springer. DOI 10.1007/s10579-009-9107-y.

Specia, L.; Srinivasan, A.; Ramakrishnan, G.; Joshi, S.; Nunes, M.G.V 2009b. An Investigation into Feature Construction to Assist Word Sense Disambiguation. *Machine Learning,* 76(1):109-136, Springer.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge applied linguistics series.

Teufel, S; Moens, M. 2002: Summarizing Scientific Articles -- Experiments with Relevance and Rhetorical Status. In *Computational Linguistics, 28* (4), Dec. 2002.

Tomazela, E.K. e Rino, L.H.M. 2009. O uso de informações semânticas para tratar a informatividade de sumários automáticos com foco na clareza referencial. Em *Anais do VII Encontro Nacional de Inteligência Artificial*, pp. 799-808. Bento Gonçalves/RS, Brasil.

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. 2008. Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory. In the *IEEE Proceedings of the 8th International Conference on Intelligent Systems Design and Applications* - ISDA, pp. 389-394. Taiwan. November, 26-28.

Watanabe, W. Candido Jr. A., Uzêda, V. Fortes, R., Pardo, T. and Aluísio, S. 2009. Facilita: reading assistance for low-literacy readers. In: *Proceedings of the 27th ACM International Conference on Design of Communication. SIGDOC '09*. ACM, New York, NY, 29-36.

Watanabe, W. M.; Candido Jr. A.; Amancio, M. A.; Oliveira, M.; Pardo, T. A. S.; Fortes, R. P. M.; Aluísio, S. M. 2010. Adapting web content for low-literacy readers by using lexical elaboration *and* named entities labeling. *Accepted for publication at W4A* 2010 (http://www.w4a.info/).

Weissberg, R.; Buker, S. 1990. *Writing up Research: Experimental Research Report Writing for Students of English.* Prentice Hall.