Uma introdução à Engenharia do Conhecimento Linguístico

Ariani Di Felippo (UNESP/Ar)
B. C. Dias-da-Silva (UNESP/Ar)

RESUMO: Neste trabalho, busca-se fornecer uma introdução ao domínio de pesquisa multidisciplinar denominado Processamento Automático de Línguas Naturais (PLN), cujas pesquisas visam capacitar os computadores a lidar com as línguas naturais. Tal domínio é definido como uma espécie de "engenharia do conhecimento linguístico". Dessa forma, o PLN requer a descrição e formalização de dados linguísticos nas dimensões morfológica, sintática, semântico-conceitual e até mesmo pragmático-discursiva. Além da concepção linguisticamente motivada do PLN, destacam-se, ao longo deste artigo, os seguintes temas: o surgimento dessa área, o lugar que ela ocupa em meio a suas disciplinas correlatas, os seus objetivos e desafios.

Palavras-chave: Processamento Automático de Línguas Naturais, Engenharia da Linguagem Humana, Linguística Computacional, Linguística, língua natural.

***ABSTRACT**: In this paper, we provide a brief introduction to the multidisciplinary domain of research called Natural Language Processing (NLP), which aims at enabling the computer to deal with natural languages. In accordance with this description, NLP is conceived as "human language engineering or technology". Therefore, NLP requires consistent description of linguistic facts on every linguistic level: morphological, syntactic, semantic, and even the level of pragmatics and discourse. Besides the linguistically-motivated conception of NLP, we emphasize the origin of such research area, the place occupied by NLP inside a multidisciplinary scenario, their objectives and challenges.*

***Keywords**: Natural Language Processing; Human Language Technology; Computational Linguistics; Linguistics; natural language.*

1. Introdução

Os computadores estão cada vez mais presentes em nosso cotidiano, como na declaração do imposto de renda ou mesmo nos caixas eletrônicos dos bancos. Essa presença massiva leva qualquer um, hoje em dia, a vincular a pesquisa científica e o desenvolvimento da tecnologia à Informática. Tal vinculação, no entanto, é menos reconhecida quando se trata de áreas consideradas menos tecnológicas, como as Ciências Humanas e Letras, como bem salienta Berber Sardinha (2005). Como consequência, não se pensa que em muitas atividades do cotidiano estão presentes tecnologias que advêm, em particular, da pesquisa sobre a linguagem com vistas ao seu processamento computacional. Tais pesquisas já haviam sido anunciadas no âmbito da Linguística, por exemplo, pela saudosa Maria Teresa Biderman, em seu texto intitulado *Teoria Linguística (linguística quantitativa e computacional*) de 1978. Atualmente, a Linguística e a Informática

encontram-se unidas em uma área de pesquisa cada vez mais promissora, denominada Processamento Automático de Línguas Naturais (PLN). As pesquisas nessa área ao mesmo tempo em que se beneficiam com os estudos provenientes da Linguística têm propiciado não só desenvolvimento de tecnologias ou recursos aplicáveis a várias atividades, mas também o próprio desenvolvimento da Linguística e da Ciência da Computação, duas das várias disciplinas matrizes do PLN.

Este texto pretende introduzir o leitor a essa área de pesquisa, enfatizando sua ligação com o estudo da linguagem. De acordo com uma concepção linguisticamente motivada do PLN, este trabalho divide-se em 6 seções. Na Seção 2, busca-se tratar as origens dessa área. Na Seção 3, discorre-se sobre a questão do lugar ocupado pelas pesquisas em PLN frente a suas disciplinas matrizes e a sua origem. Na Seção 4, destacam-se os objetivos dessa área, enfatizando as várias tecnologias por ela desenvolvidas. Na Seção 5, apresenta-se o PLN enquanto "engenharia do conhecimento linguístico" e tecem-se breves comentários sobre o processamento automático do português do Brasil. Por fim, na Seção 6, algumas considerações finais são apresentadas.

2. Eis que surge o PLN

Os computadores foram introduzidos em nossa cultura na década de 1940 e, desde então, fazê-los "entender" instruções necessárias para a execução de tarefas tem sido um desafio para os "engenheiros da linguagem". A primeira solução encontrada foi a criação das "linguagens de programação" (do inglês, *programming languages*). Com o tempo, linguagens cada vez mais inteligíveis foram criadas, como LISP, PROLOG, etc. Mesmo assim, instruções nessas linguagens são inevitavelmente rígidas, pois precisam ser descritas exatamente como o previsto.

Com a introdução dos primeiros computadores pessoais, que começaram a fazer sua história na década de 1970, a questão da comunicação entre o homem e a máquina ganhou ainda mais importância. Desenvolveu-se, como consequência, o conceito *user-friendly* – ou seja, "amigável" ou "fácil de aprender e usar" (MICROSOFT PRESS, 1998, p. 742). Esse conceito revelava a preocupação dos engenheiros da linguagem em fazer dos computadores instrumentos cada vez mais amigáveis, já que eles passavam a ser utilizados por pessoas comuns, isto é, não-especialistas. Mais especificamente, esses engenheiros buscavam tornar a comunicação entre o homem e a máquina mais natural e intuitiva, pois, a partir do momento em que a maioria dos usuários definitivamente deixava de ser

especialista em Informática, os problemas de comunicação e de significação se tornam mais importantes.

Assim, os engenheiros da linguagem passaram a pensar em possíveis linguagens que pudessem intermediar uma comunicação mais amistosa entre os computadores e seus usuários comuns. Uma das soluções encontradas (e que atualmente está presente em todos os computadores), pautada na utilização da linguagem visual, foram as chamadas "interfaces gráficas com o usuário" (do inglês, *graphical user interfaces* - GUIs), ou apenas, "interfaces gráficas". Nessa linguagem icônica, programas, arquivos e opções são representados por meio de imagens e objetos gráficos como menus, janelas, caixas de diálogos, etc. O usuário pode selecionar e ativar essas opções com o *mouse* ou, em geral, através do teclado (MICROSOFT PRESS, 1998, p. 386). Outra possibilidade seria a utilização/ adaptação da linguagem humana, ou seja, a criação de programas que pudessem, ainda que de modo rudimentar, emular o conhecimento e o desempenho linguísticos humanos. Em outras palavras, ensinar o computador a falar a língua dos homens[1] (DIAS-DA-SILVA et al., 2007).

Segundo Dias-da-Silva (2006), a possibilidade de interação homem/máquina por meio da língua dos homens e o surgimento dos primeiros sistemas de tradução automática impulsionaram os estudos ou investigações que receberam o nome *Processamento Automático de Línguas Naturais* (do inglês, *Automatic Natural Language Processing* ou *Natural Language Processing*).

3. O lugar do PLN em meio a suas disciplinas matrizes

De modo geral, no PLN, buscam-se soluções para questões computacionais que requerem o tratamento computacional de uma (ou mais) língua natural (português, inglês, etc.), quer seja escrita ou falada[2,3]. Entretanto, o processamento computacional da fala, ou melhor, das línguas naturais em modo oral, tem ficado a cargo de uma outra área, denominada Reconhecimento e Síntese de Fala (do inglês, *Speech Recognition and Synthesis*) (JURAFSKY, MARTIN, 2000). Esta, por questões tecnológicas, tem sido

[1] Na era pré-computador pessoal, a possibilidade do uso das línguas naturais na comunicação com a máquina já estava entre as questões sob investigação. No entanto, os engenheiros da linguagem visavam apenas à simplificação da vida dos programadores e técnicos que lidavam diretamente com os computadores, sem atentarem para as necessidades do usuário comum.

[2] Linguagens alternativas (p.ex.: a de sinais, para os deficientes auditivos) têm sido igualmente alvo de estudos que visam a sua automatização.

investigada pela Engenharia Elétrica, mais precisamente, na área de Processamento de Sinais. Assim, ressalta-se – e será importante dizê-lo agora – que o termo PLN aplica-se ao processamento computacional de língua natural, tanto no modo escrito quanto oral, "registrada em meio escrito".

Mais precisamente, o PLN dedica-se a investigar, propor e desenvolver sistemas computacionais que têm a língua natural escrita como objeto primário (GRISHMAN, 1986). Para tanto, os pesquisadores – linguistas e cientistas da computação – buscam fundamentos em várias disciplinas matrizes: Filosofia da Linguagem, Psicologia, Lógica, Inteligência Artificial, Matemática, Ciência da Computação, Linguística Computacional (doravante, LC) e Linguística (DIAS-DA-SILVA, 1996). No geral, em PLN, os linguistas trabalham em duas frentes: (i) utilizam o computador para desenvolver e validar teorias e dados linguísticos e (ii) fornecem o conhecimento necessário para o desenvolvimento de sistemas especializados. Os cientistas da computação, por sua vez, (i) implementam ferramentas para o desenvolvimento e validação de teorias e dados linguísticos, auxiliando os linguistas, e (ii) desenvolvem sistemas com base no conhecimento fornecido pelos linguistas.

Vê-se, assim, que o PLN é um domínio duplamente heterogêneo. O primeiro aspecto dessa heterogeneidade está ligado aos objetivos, que vão desde a proposição e desenvolvimento de programas que auxiliam a investigar material linguístico (p.ex.: programas que calculam a frequência de ocorrências de palavras em textos) até a meta de criar supercomputadores, dotados de uma inteligência artificial (JURAFSKY, MARTIN, 2000; DIAS-DA-SILVA, 2006). O segundo aspecto heterogêneo está ligado ao fato de que, para concretizar a pluralidade de objetivos, os pesquisadores necessitam percorrer as várias disciplinas matrizes, o que caracteriza esse domínio como multidisciplinar.

O objetivo abrangente e principalmente o caráter multidisciplinar do PLN dificultam delimitar o seu lugar dentre as várias disciplinas correlatas. Para as Ciências da Computação, por exemplo, o PLN é visto como uma subárea da Inteligência Artificial[4]. Isso se deve ao fato de as primeiras indagações sobre o processamento automático das línguas naturais terem sido motivadas por uma das preocupações da Inteligência Artificial, a saber: a interação homem-máquina via "língua dos homens". Muitas vezes, PLN também

[3] Como salienta Nugues (2006), o processamento de língua (escrita) e o processamento de fala são, por vezes, considerados "processamento de língua natural". Isso acontece, segundo o autor, sob o ponto de vista aplicado ou industrial.

[4] A disciplina Inteligência Artificial passou a ser reconhecida pela comunidade científica a partir da chamada *Dartmouth Summer Research Project on Artificial Intelligence* em 1956 (DIAS-DA-SILVA, 2006).

é usado como sinônimo de Linguística Computacional. Aliás, "linguística computacional" comumente nomeia grandes conferências e revistas internacionais que abrangem os estudos de PLN[5]. Entretanto, a Linguística Computacional[6], segundo Klavans (1989), Kay (1985) e outros, é o domínio que investiga questões bastante específicas do PLN, a saber: os algoritmos para as análises morfológica e gramatical. Alguns autores, por sua vez, como Bolshakov e Gelbukh (2004), consideram o PLN uma área "mais linguística que computacional" e, consequentemente, uma subárea da Linguística Aplicada. Já outros, como Nugues (2006), veem-no como uma legítima interseção entre a Linguística e as Ciências da Computação.

Além da dificuldade de delimitar o lugar desse campo dentre as disciplinas correlatas, muitos enfatizam que o corpo de conhecimento do PLN é controverso e fragmentado ou, em outras palavras, um conjunto de experiências acumuladas. Na verdade, o PLN não pertence a esta ou àquela área do conhecimento, ele é, como bem salienta Dias-da-Silva (2006), uma área de investigação científica complexa e multifacetada por natureza, sobrepondo-se, por conseguinte, a parcelas das várias áreas correlatas e já consagradas.

4. As metas

Nestes 50 anos de pesquisas, o PLN tem demonstrado ser um campo fértil em que os pesquisadores têm conseguido desenvolver tecnologias ou aplicações, com graus diferentes de sofisticação e de níveis de desempenho.

4.1 Algumas aplicações

Dentre essas tecnologias, destacam-se (BOLSHAKOV, GELBUKH, 2004; MARTINS, 2004; MITKOV, 2004; NUNES, 2008):

a) *dicionários*, *thesaurus* e *enciclopédias eletrônicas*; essas obras lexicográficas são geralmente compiladas por lexicógrafos e concebidas para uso humano, sendo armazenadas e comercializadas em CD-ROM. A microestrutura dessas obras é, na essência, a dos dicionários impressos. O fato de serem armazenados em formato digital

[5] Autores como Klavans (1989), Bolshakov e Gelbukh (2004), Mitkov (2004) e vários outros adotam a dominação Linguística Computacional e não PLN.

[6] A Linguística Computacional, cuja denominação foi cunhada por David Hays em 1967, tinha, em seus primeiros anos, o objetivo de investigar as linguagens de programação e as linguagens formais(DIAS-DA-SILVA, 2006).

contribui para potencializar toda uma rede de relações morfológicas, sintagmáticas, semânticas e paradigmáticas entre diferentes unidades lexicais e possibilitar, consequentemente, o acesso imediato à informação por outras vias que não apenas a entrada – único meio para a sua localização nos dicionários impressos (BELIAEVA et al., 1990);

b) *sistemas de recuperação de informação* (do inglês, *automatic information retrieval systems*), que buscam ou encontram textos (ou parte de textos) relevantes a uma dada "consulta" (do inglês, *query*) em uma coleção de textos ou documentos (TZOUKERMAN et al, 2004); nesses sistemas, documentos representam um tipo de informação, cuja recuperação, em outras palavras, pode ser definida como a seleção de documentos, caracterizados por um conjunto de descritores (palavras-chave ou outros símbolos), como resposta a uma consulta;

c) *sistemas de extração de informação* (do inglês, *information extraction systems*), que buscam encontrar certa informação, ou seja, uma resposta, a dada pergunta de entrada em um ou mais documentos (GRISHMAN, 2004);

d) *sistemas de tradução automática* (do inglês, *automatic translation systems*), que partem de um texto-fonte, escrito em uma língua natural *x*, e produzem um texto-alvo, ou seja, uma versão do texto-fonte em uma língua *y* (NIRENBURG, 1989; HUTCHINS, 2004; SOMERS, 2004);

e) *sistemas de sumarização automática* (SA) (do inglês, *automatic sumarization systems*): esses sistemas caracterizam-se por gerar "extratos" (justaposição de porções do texto fonte) ou "resumos" (texto gerado a partir de um plano de resumo) de um ou mais textos de acordo, por exemplo, com uma determinada taxa de compressão (HOVY, 2004).

f) *sistemas de correção ortográfica* (do inglês, *spelling checker systems*): processam um texto em uma dada língua natural com os objetivos de (i) identificar os erros cometidos quanto à ortografia (palavras que não constam do léxico dessa língua ou usadas em contexto impróprio) e (ii) sugerir alternativas prováveis e ortograficamente corretas a cada erro identificado;

g) *sistemas de correção gramatical* (do inglês, *grammar checker systems*): detectam, embora de modo rudimentar, desvios gramaticais em um texto, como os de concordância nominal ou verbal, pontuação, regência nominal e outros;

h) *sistemas de auxílio à escrita* (do inglês, *computer-assisted writing system*): auxiliam a produção de texto, em que o usuário pode encontrar recursos para construir textos bem estruturados, de um gênero e/ou domínio específicos, entre outros.

A construção dessas tecnologias nem sempre é o foco das investigações. Muitas vezes, busca-se pesquisar questões relativas a processos, métodos e recursos necessários à construção dos sistemas de PLN.

Quanto aos processos, Mitkov (2004) salienta, por exemplo: a etiquetação morfossintática (do inglês, *part-of-speech tagging),* a segmentação textual (do inglês, *text segmentation),* análise sintática (do inglês, *parsing),* a resolução da anáfora (do inglês, *anaphora resolution),* a desambiguação de sentido lexical (do inglês, *word-sense desambiguation)*, entre outros. Por vezes, a investigação desses processos resulta na construção de *ferramentas* (ou instrumentos) de PLN. Por exemplo, a investigação das questões relacionadas à etiquetação morfossintática pode levar à construção de um etiquetador morfossintático[7] (do inglês, *part-of-speech tagger*) e a investigação dos problemas relativos à análise sintática automática pode gerar um analisador sintático[8] (do inglês, *parser*). Além dessas ferramentas, há também as seguintes: lematizador[9] (do inglês, *lemmatizer*), radicalizador[10] (do inglês, *stemmer*), concordanceador (do inglês, *concordancer*), entre outras. Algumas delas são componentes essenciais de vários sistemas.

Quanto aos métodos ou técnicas, os pesquisadores têm investigado a viabilidade de diferentes abordagens para a construção de sistemas de PLN. Atualmente, co-existem pesquisas realizadas segundo abordagens linguísticas, não-linguísticas (ou estatísticos) e híbridas (MARTINS, 2004). As abordagens linguísticas pautam-se na especificação explícita e declarativa de propriedades e de regras ou padrões regulares de comportamento linguístico. As abordagens não-linguísticas, por sua vez, pautam-se na recuperação/identificação, induzida automaticamente, de regularidades subjacentes aos dados linguísticos e, por isso, necessitam de extensos *corpora* para que os padrões possam

7 Ferramenta computacional responsável pela marcação de um texto com etiquetas morfossintáticas. Esses etiquetadores podem ser construídos manualmente, por linguistas, ou automaticamente, abstraídos de *corpus*. (VOUTILAINEN, 2004).

8 Ferramenta que reconhece a estrutura sintática de uma sentença, atribuindo funções sintáticas aos constituintes reconhecidos (CARROL, 2004).

9 Ferramentas que reduzem cada palavra de um texto ao seu lema ou forma canônica, ou seja, formas não-marcadas, desprovidas de flexões (SPARCK-JONES, WILLET, 1997). Na lematização, os verbos são reduzidos ao *infinito* (p.ex.: casamos > casar) e os substantivos e adjetivos ao *masculino singular* (p.ex.: latas > lata/ feias > feio).

ser identificados. As estratégias híbridas, por fim, reúnem as características das linguísticas e das não-linguísticas (DORR et al, 1999).

4.2 Os recursos ou ferramentas

A construção de certas ferramentas e a aplicação de determinados métodos ou técnicas, ambos importantes para o subsequente desenvolvimento de sistemas de PLN, necessitam, quase sempre, dos chamados *recursos linguístico-computacionais*, cujo planejamento (e construção) constitui tarefa nada trivial. Exemplos desses recursos são:

a) *corpora* (textuais): coleções de textos úteis para o levantamento de conhecimento linguístico (lexical, sintático, semântico, etc.). Esse levantamento pode ser feito por linguistas, com a ajuda de programas de manipulação de *corpus*, ou por meio da aplicação de métodos estatísticos. A extração do conhecimento exige que a quantidade de textos seja grande, variada e representativa e que os textos estejam em formato adequado para que a extração possa ser automática (BERBER SARDINHA, 2004);

b) *léxicos*: estoques de unidades lexicais descritas juntamente com seus traços morfológicos, sintáticos, semânticos e/ou pragmático-discursivos e sistematicamente organizadas de acordo com algum critério. Tanto as unidades quanto as propriedades a elas associadas podem ser representadas por formalismos altamente estruturados (HANDKE, 1995);

c) *ontologias* e/ou *bases de conhecimento*: inventários de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade", ou seja, o conhecimento de mundo compartilhado pelos membros de uma comunidade linguística. A representação de uma ontologia pode variar segundo o grau de formalização. Uma ontologia formal, em especial, apresenta os conceitos e as relações (entre conceitos) explicitamente definidas, ou seja, "legíveis pela máquina" (GRUBER, 1995);

d) *gramáticas:* sistemas de regras expressos segundo sistemas formais, que (i) descrevem as estruturas das sentenças de uma dada língua e (ii) permitem, juntamente com o léxico, reconhecer e gerar sentenças dessa língua (KAPLAN, 2004).

Conclui-se, assim, que o PLN é uma área complexa e multifacetada e que, mesmo aparentemente caótica, tem se mostrado produtiva. Por fim, salienta-se que o PLN também

[10] Ferramentas computacionais que reduzem as palavras de um texto ao seu radical (SPARCK-JONES, WILLET, 1997).

possui um viés acadêmico e não somente científico-tecnológico. Dentre os objetivos dos pesquisadores, estão (i) a investigação da adequação formal, pragmática e psicossocial de teorias linguísticas por meio da implementação dos modelos de gramática e de processamento linguístico especificados por essas teorias e a própria (ii) proposição de sofisticados modelos computacionais capazes, por exemplo, de extrair informações específicas de bases de textos (VARELI, ZAMPOLLI, 1997).

5. A face linguística do PLN

Segundo Dias-da-Silva (2006), que se baseia em Winograd (1972), um sistema de PLN pode ser visto como um tipo especial de "sistema especialista" na medida em que requerem uma parcela específica do conhecimento humano – o conhecimento linguístico – para realizar tarefas específicas como correção ortográfica, tradução automática, etc.

No âmbito da Inteligência Artificial, um *sistema especialista* (do inglês, *expert system*) é um sistema computacional inteligente, que toma decisões e resolve problemas referentes a um determinado campo de atuação, como finanças e medicina, utilizando conhecimento e regras analíticas definidas por especialistas no assunto (JACKSON, 1990; HAYES-ROTH, 1990; MICROSOFT PRESS, 1998; GIARRATAMO, RILEY, 2004). Um sistema de diagnóstico, por exemplo, necessita saber quais as características das doenças a serem diagnosticadas, pois, sem elas, é impossível elaborar um diagnóstico automaticamente. Dentre os sistemas especialistas descritos na literatura, destacam-se o (i) Dendral, primeiro sistema especialista, criado para ajudar os químicos a determinar a estrutura molecular, (ii) o Mycin, que diagnostica doenças sanguíneas infecciosas, e o (iii) Dipmeter Advisor, que auxilia na análise de dados recolhidos durante a exploração de óleo.

Projetar, então, um sistema de PLN, ou seja, um sistema que simule parcelas da competência e do desempenho linguístico humanos, requer a especificação de vários conhecimentos e habilidades que os falantes (especialistas nesse domínio) possuem. Esse embasamento linguístico fica evidente nas palavras de Winograd:

> Assumimos que um computador não poderá simular uma língua natural satisfatoriamente se não compreender o assunto que está em discussão. Logo, é preciso fornecer ao programa um modelo detalhado do domínio específico do discurso. Além disso, o sistema possui um modelo simples de sua própria mentalidade. Ele pode se lembrar de seus planos e ações, discuti-los e executá-los. Ele participa de um diálogo, respondendo, com ações e frases, às frases digitadas em inglês pelo usuário; solicita esclarecimentos quando seus programas heurísticos não conseguem compreender uma frase com a ajuda das informações sintáticas,

> semânticas, contextuais e do conhecimento de mundo físico representados dentro do sistema. (WINOGRAD, 1972, tradução nossa).

Mais precisamente, acredita-se que, para simular uma língua natural de modo satisfatório, um sistema de PLN deve conter vários sistemas de "conhecimento" e "realizar" uma série de atividades cognitivas, tais como (DIAS-DA-SILVA et al, 2007): (i) possuir um "modelo simples de sua própria mentalidade"; (ii) possuir um "modelo detalhado do domínio específico do discurso"; (iii) possuir um modelo que represente "informações morfológicas, sintáticas, semânticas, (iv) contextuais e do conhecimento de mundo físico"; (v) "compreender o assunto que está em discussão"; (vi) "lembrar, discutir, executar seus planos e ações"; (vii) participar de um diálogo e responder, com ações e frases, às frases digitadas pelo usuário; (viii) solicitar esclarecimentos quando seus programas heurísticos não conseguirem; (ix) compreender uma frase

Dessa forma, um sistema de PLN é concebido como um tipo de sistema automático de conhecimentos, cujas especialidades, entre outras, incluem: fazer revisões ortográficas de textos, fazer análises sintáticas, traduzir frases ou textos, fazer perguntas e respostas e auxiliar os pesquisadores na própria construção de modelos linguísticos. Assim, o estudo do PLN pode ser concebido como um tipo de "engenharia do conhecimento linguístico" e se beneficiar da estratégia desenvolvida para esse campo.

Assim, de modo semelhante ao processo de construção de um sistema especialista, também denominado "sistema de conhecimento" (do inglês, *knowledge system*), a montagem de um sistema de PLN exige o desenvolvimento de, no mínimo, três etapas: "extração do solo" (explicitação dos conhecimentos e habilidades linguísticas), "lapidação" (representação formal desses conhecimentos e habilidades) e "incrustação" (o programa de computador que codifica essa representação). Assim, a explicitação do conhecimento e do uso linguísticos envolve questões do domínio linguístico, uma vez que é nessa fase que os fatos da língua e do seu uso são especificados. Conceitos, termos, regras, princípios, estratégias de resolução de problemas e formalismos linguísticos são os elementos trabalhados. No domínio da representação, focalizam-se questões referentes à escolha ou à proposição de sistemas de representação. No domínio da implementação, além das questões que envolvem a implementação das representações por meio de programas, há questões que dizem respeito à montagem do próprio sistema computacional em que o programa será alojado.

As três etapas previstas para o desenvolvimento de sistemas especialistas foram reinterpretadas por Dias-da-Silva (1996) e transformadas em uma metodologia que vem sendo aplicada com sucesso no âmbito do PLN, principalmente no que se refere ao processamento automático do português do Brasil. Tal metodologia consiste em três fases sucessivas de desenvolvimento das atividades no PLN, a saber: *fase linguística:* construção do corpo de conhecimentos sobre a própria linguagem, dissecando e compreendendo os fenômenos linguísticos necessários para o desenvolvimento do sistema; *fase representacional:* construção conceitual do sistema, envolvendo a seleção e/ou proposição de sistemas formais de representação para os resultados propostos pela fase anterior; *fase implementacional*: codificação das representações elaboradas durante a fase anterior em termos de linguagens de programação e planejamento global do sistema.

Para as pesquisas que adotam a "concepção linguisticamente motivada de PLN", o computador não poderá satisfatoriamente emular uma língua natural se não conseguir, em alguma medida, compreender o assunto que está em discussão. Logo, é preciso fornecer à máquina descrições e formalizações de dados linguísticos nas dimensões: morfológica, sintática, semântico-conceitual e pragmático-discursiva (ROCA, 2000). E aí a Linguística tem um papel imprescindível, pois, apesar dos aspectos problemáticos comumente apontados pelos engenheiros da linguagem, ela apresenta os parâmetros norteadores essenciais a respeito das características e funções das línguas naturais a que os investigadores do PLN podem recorrer.

Para o desenvolvimento de uma pesquisa "linguisticamente motivada de PLN", acredita-se ser necessário, como defende Dias-da-Silva (1996, 1998, 2006), o trabalho colaborativo entre os cientistas e os engenheiros da linguagem.

Essa colaboração, entretanto, está longe de ser a ideal. Há, ainda, o distanciamento entre essas duas comunidades, o que dificulta e/ou atrasa a descoberta de soluções e o consequente avanço no desenvolvimento dos recursos, ferramentas e, logo, dos sistemas (DIAS-DA-SILVA, 1996). Tal distanciamento tem sido justificado por razões técnicas fornecidas por ambos os lados. Os engenheiros criticam, por exemplo, a pluralidade, a incompletude e a pouca formalização das descrições linguísticas, o linguajar técnico muitas vezes hermético e a preocupação dos linguistas em estudar a linguagem humana *per se.* Os linguistas, por sua vez, enfatizam que os engenheiros – tidos como indivíduos com pouca intuição sobre os fatos da língua – concentram-se no desenvolvimento de sistemas

rudimentares e desprovidos de qualquer fundamentação linguística[11]. A falta de contato entre essas duas comunidades, aliás, também é regada por imagens estereotipadas e distorcidas que os pesquisadores de uma área têm do trabalho realizado na outra, principalmente quando as áreas de conhecimento são tão distintas, como Linguística e Ciências da Computação.

Nos casos em que o distanciamento foi vencido, a colaboração entre linguistas e cientistas da computação mostrou-se não somente benéfica para o PLN, mas também para a Linguística e Ciência da Computação. A Linguística, por exemplo, tem se beneficiado, do ponto de vista prático, com vários recursos que auxiliam na análise de material linguístico. Do ponto de vista teórico, tem se beneficiado também com a formulação de modelos descritivos mais completos (ou seja, modelos de análise e descrição de cada um dos estratos da gramática e do inter-relacionamento entre os módulos da competência e do desempenho) e explícitos (ou seja, descritos em termos de linguagens formais). Isso acontece porque, como evidenciou Winograd, pesquisar o PLN pode ser também um modo de investigação acadêmico que pode auxiliar na compreensão dos próprios fatos da língua:

> Todo mundo é capaz de compreender uma língua. A maior parte do tempo de nossas vidas é preenchida por atos de fala, leitura ou pensamentos, sem sequer notarmos a grande complexidade da linguagem. Ainda não sabemos como nós sabemos tanto […] Os modelos [de PLN] são necessariamente incompletos […] Mas, mesmo assim, constituem um referencial claro por meio do qual podemos refletir sobre o que é que fazemos quando compreendemos uma língua natural ou reagimos aos atos de fala nela codificados. (WINOGRAD, 1972, tradução nossa).

Ou ainda, nas palavras de Hoey (*apud* Berber Sardinha, 2005, p. 30):

> O desenvolvimento do computador com memória poderosa seria para a linguística o que a desenvolvimento do microscópio com lentes poderosas foi para a biologia – uma oportunidade não somente de ampliar nosso conhecimento, mas de transformá-lo.

Alguns exemplos paradigmáticos da contribuição do PLN para os estudos linguísticos são: (i) implementação, teste e avaliação de gramáticas propostas pela Linguística Teórica (GRISHMAN, 1986), como a gramática funcional de Dik (1997) (SIEWIERSKA, 1991; ATKINS, ZAMPOLLI, 1994) e parcela da gramática funcional de Halliday (1985) (BUTLER, 1985); (ii) desenvolvimento de modelos gramaticais, como a

[11] Por exemplo, os próprios dicionários eletrônicos, em que o material linguístico é apenas manipulado por meio de técnicas de indexação, e os chamados "tradutores de bolso", que são limitados a manipular listas de palavras, expressões e fragmentos de frases por meio de comparações e substituições com o objetivo de montar/ completar frases com as palavras e/ou expressões armazenadas (DIAS-DA-SILVA, 2006).

HPSG (POLLARD, SAG, 1994); (ii) proposição de modelos diversos, p.ex.: modelos computacionais dos atos de fala, modelos computacionais da teoria da referência (DIAS-DA-SILVA, 2006).

Vale ressaltar que, por aspectos econômicos, as tecnologias em PLN são, na maioria das vezes, desenvolvidas para a língua inglesa, sendo que tais tecnologias não são diretamente transportáveis para outras línguas. Dessa forma, o processamento de uma língua natural requer o desenvolvimento de recursos e ferramentas de base que deem suporte para o desenvolvimento de sistemas voltados para a língua em questão.

6. Considerações finais

O PLN, entendido como uma espécie de "engenharia do conhecimento linguístico", é um campo de pesquisa privilegiado e fértil. Isso é reflexo do fato de que a delimitação do conhecimento necessário para a construção de sistemas de PLN exige a organização e a representação de uma variedade de dados complexos necessários à simulação da competência e do desempenho linguísticos (DIAS-DA-SILVA, 2006). Devido a essa exigência, tanto usuários leigos, por meio da construção de várias aplicações, quanto os próprios linguistas e cientistas da computação têm se beneficiado com as investigações no campo do PLN. Dessa forma, encerra-se este artigo com o mote "cooperar é preciso", enfatizado a relevância do trabalho colaborativo entre os cientistas da linguagem ou linguistas e os engenheiros da linguagem.

Referências Bibliográficas

ATKINS, S, ZAMPOLLI, A. **Computational approaches to the lexicon**. Oxford: Oxford University Press, 1994.

BELIAEVA, L. N., PIOTROWSKI, R. G., SOKOLOVA, S. V. *Principles of linguistic automata and their information bases design. In* **Terminology and Knowledge Engineering**, v. 2, p. 419-425, 1990.

BERBER SARDINHA, A. P. **Linguística de corpus**. São Paulo, Barueri: Editora Manole, 2004.

___. (Org.). **A língua portuguesa no computador**. Campinas, SP: Mercado de Letras; São Paulo: FAPESP, 2005.

BIDERMAN, M.T.C. **Teoria linguística** (linguística quantitativa e computacional). Rio de Janeiro/São Paulo, LTC, 1978.

BOLSHAKOV, I., GELBUKH, A. **Computational Linguistics:** models, resources and applications. México City: Centro de Investigación en Computación/ Instituto Politécnico Nacional, 2004.

BUTLER, C. S. **Systemic linguistics:** theory and applications. London: Batsford Academic and Educational, 1985.

CARROL, J. Parsing. *In*: MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 12, p. 233-248.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem**: o processamento automático das línguas naturais. Araraquara, 1996. 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.

_____. *Bridging the gap between linguistic theory and natural language processing*. *In*: INTERNATIONAL CONGRESS OF LINGUISTICS, 16, 1997, Paris. **Proceedings…** Oxford: Elsevier Sciences, 1998, n. 16, p. 1-10.

_____. *O estudo linguístico-computacional da linguagem*. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

_____ et al. Introdução ao Processamento das Línguas Naturais e algumas aplicações. **Série de Relatórios Técnicos do NILC**, NILC-TR-07-10. São Carlos, 2007, 121p.

DIK, S. C. **The theory of functional grammar**. Berlin, New York: Mouton de Gruyter, 1997.

DORR, B.J. JORDAN, P.W., BENOIT, J.W. *A survey of current research in machine translation. In* **Advances in Computers**, v. 49, p. 1-68, 1999.

GIARRATANO, J.C., RILEY, G.D. **Expert systems***: principles and programming*. Boston: Course Techonology, 2004.

GRISHMAN, R. **Computational linguistics**. Cambridge: Cambridge University Press, 1986.

_____. *Information extraction*. *In*: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 30, p. 545-559.

GRUBER, T. *Toward principles for the design of ontologies used for knowledge sharing*. **International Journal Human-Computer Studies**, v. 43, n. 5-6, p. 907-928, 1995.

HALLIDAY, M.A.K. **An introduction to functional grammar**, London: Edward Arnold, 1985.

HANDKE, J. **The structure of the lexicon: human vs machine**. Berlin: Mouton de Gruyter, 1995.

HAYES-ROTH, F. Expert systems. In: SHAPIRO, E. (Ed.). **Encyclopedia of artificial intelligence**. New York, Wiley, 1990, p. 287-298.

HOEY, M. **Patterns of lexis in text**. Oxford: Oxford University Press, 1991.

HOVY, E. *Text summarization*. *In*: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics.** Oxford: Oxford University Press, 2004, cap. 32, p. 583-598.

HUTCHINS, W.J. *Information extraction*. *In*: MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford: Oxford University Press, 2004, cap. 30, p. 545-559.

JACKSON, P. **Introduction to expert systems**. Wokingham: Addison-Wesley, 1990.

JURAFSKY, D., MARTIN, J.H. **Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition**. Upper Saddle River, New Jersey: Prentice Hall, 2000.

KAPLAN, R. M. *Syntax*. *In*: MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford: Oxford University Press, 2004, cap. 04, p. 70-90.

KAY, M. *Parsing in functional unification grammar*. *In*: DOWTY, D. R. et al. (Eds.). **Natural language parsing**. Cambridge: CUP, 1985, p. 251-278.

KLAVANS, J. *Computational linguistics*. *In*: O'GRADY, W. et al. **Contemporary linguistics**. New York: St. Martin's Press, 1989, cap. 15, p. 413-447.

MARTINS, R.T. **A nova língua do imperador**. Campinas, 2004. 296p. Tese (Doutorado) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2004.

MICROSOFT PRESS. **Microsoft press dicionário de informática**. Rio de Janeiro: Editora Campus, 805 p., 1998.

MITKOV, R. (Ed.). **The Oxford handbook of Computational Linguistics**. Oxford, New York: Oxford University Express, 2004.

NIRENBURG, S. *Knowledge and choices in machine translation*. *In*: NIRENBURG, S. (Org.). **Machine translation** – theoretical and methodological issues. Cambridge: Cambridge University Press, 1989, p. 1-15.

NUGUES, P. M. (2006*).* **An introduction to language processing with perl and prolog**. Springer-Verlag, 2006.

NUNES, M.G.V. *Processamento de línguas naturais*: para quê e para quem. *In* **Notas Didáticas** *ICMC-USP,* 73, 11p, 2008.

POLLARD, C., SAG, I. **Head-driven phrase structure grammar**. Chicago: University of Chicago Press, 1994.

ROCA, S. C. *Individuación e información parte-todo:* Representación para el procesamiento computacional del lenguaje. **Estudios de Linguística Española**, v. 08, 2000. Disponível em <http://elies.rediris.es/elies8/>. Acesso em: 10 jun. 2005.

SIEWIERSKA, A. **Functional grammar**. London-New York: Routledge, 1991.

SOMERS, H. *Machine translation*: latest developments. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 28, p. 512-528.

SPARCK-JONES, K., WILLET, P. **Readings in information retrieval**. São Francisco: Morgan Kaufmann, 1997.

TZOUKERMAN, E., KLAVANS, J. L., STRZALKOWSKI, T. *Informaton retrieval*. *In*: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 29, p. 529-544.

VARELI, G. B., ZAMPOLLI, A. **Survey of the state of the art in human language technology**. Cambridge: CUP, 1997.

VOUTILAINEN, A. *Part-of-speech tagging*. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 11, p. 219-232.

WINOGRAD, T. **Understanding natural language**. New York: Academic Press, 1972.

Ariani Di Felippo é mestre em Linguística e Língua Portuguesa pela Universidade Estadual Paulista Júlio de Mesquita Filho, e doutoranda da Faculdade de Ciências e Letras da Universidade Estadual Paulista. Tem experiência na área de Linguística, com

ênfase em Linguística Computacional ou Processamento Automático das Línguas Naturais, atuando principalmente nos seguintes temas: Semântica Lexical Computacional e Lexicologia/Terminologia Computacional. (arianidf@uol.com.br)
B. C. Dias-da-Silva é doutor em Linguística e Língua Portuguesa pela Universidade Estadual Paulista Júlio de Mesquita Filho e mestre em Linguística e Língua Portuguesa pela mesma universidade. Atualmente é pesquisador da Universidade de São Paulo, pesquisador da Universidade Federal de Santa Catarina e professor assistente doutor da Universidade Estadual Paulista Júlio de Mesquita Filho. Tem experiência na área de Linguística, com ênfase em Teoria e Análise Linguística, atuando principalmente nos seguintes temas: léxico, semântica lexical, wordnets, processamento automático de línguas naturais e relações de sentido. (bento@fclar.unesp.br)