O estudo Lingüístico-Computacional da Linguagem

Bento Carlos Dias da Silva*
Universidade Estadual Paulista

1 Introdução

Desde a sua introdução na cultura ocidental no início dos anos 40 os computadores não só vêm contribuindo para avanços substantivos nos diversos campos do conhecimento científico como também têm sido responsáveis pelo desenvolvimento e pela abertura de novas frentes de pesquisa que sem eles não teriam sido cogitadas. Capazes de subsidiar a resolução rápida e precisa de uma infinidade de problemas científicos complexos, de substituir o homem em tarefas arriscadas, repetitivas e estafantes, de auxiliar o usuário comum nas tarefas cotidianas e de proporcionar lazer e entretenimento a usuários de todas as idades, essas máquinas, admiradas por uns, ignoradas por outros e, até mesmo, temidas por muitos, estão por toda parte e viabilizaram a surpreendente "revolução da informação" iniciada no final do milênio e responsável pela massa de dados que circula pela intrincada rede de computadores que, hoje, materializa-se, sobretudo, na Internet.

Dentre os grandes desafios que essas máquinas lançaram aos seus arquitetos, destaca-se o desafio de se criar meios para tornar a comunicação homem-máquina mais natural e intuitiva. Uma das soluções para esse desafio, e, hoje, implementada em todos os computadores, são as interfaces gráficas e multicoloridas que servem para mediatizar a comunicação com o computador. Essas interfaces são compostas por milhares de linhas de código de máquina capazes de codificar as instruções necessárias para fazer com que o computador realize tarefas em termos de objetos gráficos, isto é, em termos de janelas, ícones, menus, caixas de diálogo, imagens, entre

* E.mail: bento@fclar.unesp.br

outros recursos.[1] Na realidade, essas interfaces gráficas, como máscaras, escondem a única linguagem decifrável pelo computador: a linguagem de máquina que, em última instância, representa toda a informação manipulada pelo sistema computacional em termos de estruturas representadas exclusivamente por cadeias de "zeros e uns".

Uma outra solução para o mesmo desafio seria o desenvolvimento de programas capazes de "compreender", pelo menos de modo rudimentar, fragmentos da linguagem humana, alternativa cujos reflexos são sentidos até hoje, diante das inúmeras tentativas de se investigar como fazer o computador emular o conhecimento e o desempenho lingüísticos humanos. A essa preocupação com a comunicação natural, que já se instalava nos círculos universitários norte-americanos e europeus, concomitantemente com a criação dos primeiros computadores, somou-se outra não menos complexa: a iniciativa voltada para o desenvolvimento de sistemas de tradução realizada automática ou semi-automaticamente por computador.

É o desafio posto pelo tratamento computacional das línguas naturais e pelo próprio processo de comunicação humano que tem instigado os centros de tecnologia da linguagem humana a investirem significativos recursos teóricos, humanos e materiais na modelagem computacional da linguagem humana, entendida, aqui, como a criação de um modelo computacionalmente tratável do uso do léxico e da gramática de uma língua natural nas diversas situações comunicativas. Nasce, assim, o domínio de estudo conhecido por *Processamento Automático de Línguas Naturais* (doravante PLN).

Uma peculiaridade desse amplo e controverso domínio de pesquisa é o fato de agregar uma heterogeneidade de objetivos: desde a meta de investigar meios de empregar o computador como uma simples ferramenta auxiliar para investigar material lingüístico (por exemplo, a criação de programas de computador para calcular estatísticas de ocorrências de palavras em textos ou para identificar e indexar palavras e segmentos de texto) até a meta de criar uma inteligência artificial, nos moldes do supercomputador

---

[1] As plataformas gráficas passaram a ser desenvolvidas na década de 80. Precisamente em 1984, fazendo uma alusão direta a essa data, título e ícone da obra de George Orwell, os seus idealizadores, as empresas americanas *Apple*™ e *Xerox*™, passaram a comercializar o computador que ficou mundialmente conhecido como *Macintosh*™: o primeiro computador pessoal equipado com uma sofisticada interface gráfica. Estava quebrado o monopólio da informática pela gigante *IBM*™. Depois dessa iniciativa, esta e a então principiante *Microsoft*™ uniram-se e passaram também a desenvolver suas próprias plataformas gráficas concorrentes. Nasce, então, a plataforma *Windows*™, que ainda hoje, mesmo com as iniciativas de desenvolvimento de plataformas abertas, como o sistema *Unix* e a sua versão contemporânea *Linux*, domina a maioria dos computadores de todo o mundo.

HAL-9000 do clássico de Stanley Kubrick – *2001: Uma Odisséia no Espaço*. Em outras palavras, nesse domínio há uma considerável pluralidade de objetivos e interesses: desde o estudo meramente quantitativo das línguas, que, na essência, visa à construção de listas de freqüência de palavras, listas de concordâncias, lista de lemas e análise de possibilidades combinatórias de unidades lingüísticas, passando pelo estudo da adequação formal, pragmática e psicossocial de teorias lingüísticas, por meio da implementação computacional dos modelos de gramática e de processamento lingüístico por elas especificados, até a proposição de sofisticados modelos computacionais capazes de extrair informações específicas de bases de textos, de propor a sumarização ou a tradução de textos e até mesmo manter um diálogo livre com o usuário em língua natural (Varile & Zampolli, 1997).

Diante dessa pluralidade, com mais de meio século de experiências acumuladas, algumas bem-sucedidas e outras absolutamente desastrosas, o trabalho desenvolvido no domínio do PLN acumula uma vasta literatura, produzida por estudiosos das mais variadas disciplinas, com formações acadêmicas, embasamentos teóricos e interesses de pesquisa também bastante diversos. O agravante, porém, é constatar que, por razões diversas, os complexos fenômenos da linguagem, cuja compreensão é, sem dúvida, condição essencial para o sucesso de qualquer empreendimento em PLN, têm sido subestimados e, conseqüentemente, subdimensionados no processo de desenvolvimento dos mais variados tipos de sistemas computacionais que, de alguma forma, envolvem o tratamento automático de entidades e processos lingüísticos.

Essa constatação, por um lado, evidencia a escassez, senão a falta, de modelos lingüísticos computacionalmente implementáveis, o que tem deixado aos informatas a tarefa de, às duras penas, construírem seus "léxicos", "gramáticas", "analisadores gramaticais" e "planejadores de discurso"; por outro, revela o incômodo distanciamento que se estabelece entre o estudo do PLN e o estudo da linguagem humana, sobretudo porque este não reconhece no âmbito daquele um espaço de investigação legítimo para a compreensão dos fenômenos lingüísticos.

Começar a reverter esse quadro implica apontar os entraves que dificultam a aproximação entre os "cientistas" e os "engenheiros" da linguagem. Aqueles são os investigadores que se concentram na investigação das línguas naturais *per se* e estes, na modelagem computacional destas. Em um segundo momento, implica também propor um equacionamento estratégico global para a pesquisa em PLN que evidencie a importância e a necessidade de se

projetarem sistemas de PLN que sejam, no limite, computacionalmente eficientes e robustos e lingüisticamente precisos e adequados e que saliente a relevância acadêmica da investigação do PLN para o avanço quantitativo e qualitativo dos estudos contemporâneos da linguagem, ao apontar as dimensões científicas e tecnológicas inerentes ao estudo computacional das línguas naturais.

No contexto dessas considerações iniciais é que defino os contornos da face lingüístico-computacional dos estudos da linguagem, que, com será visto, delineia-se, porém de modo difuso, tanto nas teorias lingüísticas contemporâneas como no vasto e disperso domínio de estudo do PLN.

Nessa tarefa de estabelecimento de contornos, aponto, na segunda seção deste artigo, o que considero entraves para o trabalho cooperativo entre lingüistas e projetistas de sistemas de PLN. Na terceira seção, após delimitar a concepção de PLN lingüisticamente motivada, proponho o equacionamento metodológico global que considero essencial para o desenvolvimento da pesquisa nessa área. Por fim, na quarta seção, sistematizo as relações de interdisciplinaridade que se estabelecem entre a investigação do PLN e as disciplinas matrizes que lhe dão fundamentação e mapeio os recursos teóricos essenciais para o desenvolvimento dos sistemas de PLN.

2 Entraves

Tentar construir um corpo de conhecimentos suficientemente estruturados e integrados, capaz de fornecer os meios que poderão transformar máquinas em "tradutores ou interlocutores cibernéticos" é um empreendimento arrojado e fascinante. Entretanto, ao se aventurar em participar desse empreendimento, constata-se um cenário de desencontros: uma preocupante lacuna entre o estudo do PLN e a Teoria Lingüística. Starosta (1991, p. 178) sinaliza com precisão o desconforto:

> [Por um lado] a fragmentação do campo da lingüística e as especulações filosóficas nebulosas que se fazem passar por "teoria lingüística" nos grandes círculos da comunidade de lingüistas não inspiram confiança nos engenheiros da linguagem; [por outro] o total desconhecimento dos fatos elementares da linguagem que esses engenheiros ostentam presunçosamente em suas publicações em nada contribui para convencer os lingüistas sérios que, lá entre os computadores, haja alguém que esteja interessado em aplicar o que quer que seja que os cientistas da linguagem tenham descoberto.

2.1 *Desafio para os engenheiros da linguagem*

Do ponto de vista dos projetistas de sistemas de PLN, é possível encontrar razões que os levem a se distanciar dos estudos lingüísticos. Como ponto de partida, relembro parte das dificuldades que enfrentei ao desenvolver meu trabalho de mestrado (Dias-da-Silva, 1990) no domínio da Teoria Lingüística. Na busca de modelos teóricos para investigar o fenômeno da apassivação nas línguas naturais, o que encontrei foi um "frenético borbulhamento de novidades teóricas". Escolher um arcabouço descritivo para fundamentar a pesquisa tornou-se, para mim, um problema muito mais complexo do que para os pesquisadores de outras áreas do conhecimento, "que têm a felicidade de poder contar com um cerne básico estável de princípios epistemológicos e convenções notacionais".

As expressões destacadas entre aspas no parágrafo anterior e neste parágrafo, tomadas de empréstimo de Lemle (1984, p. 2), ainda traduzem parte das razões que têm distanciado o não-lingüista da Teoria Lingüística. Lemle, ao propor uma "ponte" entre a lingüística teórica e o ensino escolar da gramática, comenta que "uma certa lingüística" emprega um "formalismo algébrico bizarro, abominável e desinteressante para a maioria das pessoas" e critica a própria Teoria Lingüística, ao afirmar que o discurso lingüístico vale-se com "demasiada freqüência de um linguajar técnico hermético que disfarça o vazio de substância de suas propostas".

Esses exemplos já são suficientes para se ter uma idéia de que a escolha e a avaliação de modelos descritivos no âmbito da Lingüística transformam-se em tarefas laboriosas e desnorteadoras. Não raro, as propostas lançadas pelos teóricos trazem consigo uma pluralidade de análises, muitas delas simplesmente esboçadas, ou uma metalinguagem hermética e cifrada ou ainda uma evidente concentração em aspectos particulares e pontuais da análise das línguas. Qualquer lingüista, por exemplo, reconhece o "viés sintático" que dominou as pesquisas lingüísticas na segunda metade do século XX.

A fragmentação, a parcialidade e a pouca formalização das descrições de fatos lingüísticos são também apontadas como agravantes para o quadro de distanciamento. Winograd (1972, p. 41), um pesquisador do PLN, por exemplo, apresenta uma justificativa plausível para o fato dos engenheiros da linguagem não recorrerem à Teoria Lingüística:

> Quando surgiram os primeiros trabalhos de descrição de línguas naturais com o auxílio do computador, não havia teorias sintáticas suficientemente explícitas e prontas para serem computacionalmente codificadas. Os primeiros projetistas [de sistemas de PLN] que ousaram propor os sistemas de tradução automática pioneiros foram forçados a construir seus próprios modelos de descrição lingüística... Como decorrência, [esses modelos] eram muito precários e imediatistas.

Até muito recentemente, por ser considerada território muito complexo e difícil de ser explorado, a semântica era alvo de comentários irônicos. Hirst (1992, p. 1), comentando que os lingüistas tratavam-na como algo "excelente para se discutir, porém incognoscível", possuindo as "mesmas qualidades de Deus ou da Mente", apresenta outra justificativa para a lacuna entre os cientistas e os engenheiros da linguagem:

> Posto que havia lingüistas com atitudes como essa em relação à semântica, não é de causar surpresa que os consumidores de teorias lingüísticas, tais como os pesquisadores do PLN, tomassem, eles próprios, a iniciativa de estudarem a semântica.

Já Schank e Riesbeck (1981, p. 2) acusam os lingüistas não só de se fixarem demasiadamente nos estudos sintáticos e minimizarem os estudos semânticos como também, e sobretudo, de pouco se preocuparem com o estudo do uso da linguagem:

> Quando surgiram os primeiros trabalhos sobre PLN, os pesquisadores passaram a aplicar qualquer teoria disponível. As implementações de teorias sintáticas diversas tiveram uma certa popularidade no domínio da Inteligência Artificial durante algum tempo. Enquanto que a real preocupação nesse domínio, porém, sempre recaiu sobre o tratamento do significado, os lingüistas, durante muito tempo, evitaram abordá-lo em suas investigações.

A crítica aos lingüistas continua ecoando com Mykowiecka (1991, p. 497):

> Um dos motivos que vem impedindo o rápido desenvolvimento do estudo do PLN é o fato de que a maioria dos lingüistas não está disposta a cooperar [...], o que, conseqüentemente, acarreta uma escassez de teorias lingüísticas e de definições suficientemente precisas para o uso computacional.

Há lingüistas que chegam a criticar severamente os grupos de pesquisa envolvidos com o PLN. Moreno Fernández (1990), por exemplo, diz que esses grupos só existem para alimentar a indústria de informática: cada grupo trabalha para oferecer ao mercado

consumidor programas mais sofisticados, mais eficientes e mais econômicos que os programas desenvolvidos por seus pares. Acrescenta que, por esse motivo, é muito difícil encontrar publicações que informem com regularidade e transparência os avanços alcançados por parte das instituições, sejam elas públicas ou privadas. Em outras palavras, a informação não é divulgada para o público interessado, porque isso significaria ceder resultados para um competidor potencial.

A demanda urgente de aplicativos e a limitação de recursos computacionais são também agravantes que têm impedido que os sistemas de PLN passem a incorporar sofisticações que exijam estudos mais aprofundados e que sejam consistentes com a Teoria Lingüística. Essas são justificativas do ponto de vista comercial, mas não se sustentam do ponto vista científico e tecnológico. Allen (1987, p. 2), entretanto, já alertava para esse risco e é categórico ao afirmar que os objetivos mercadológico não poderão ser alcançados sem se buscar fundamentação nas sofisticadas teorias propostas no âmbito da lingüística teórica.

Esse distanciamento entre o PLN e os estudos da linguagem acaba por ser justificado se considerar-se que a Lingüística, com recentes e tímidas iniciativas de inclusão de trabalhos sob o rótulo genérico "Lingüística Computacional" na sua agenda de pesquisa, não é afeita a atender as demandas do PLN. Ocupados com o estudo da linguagem humana *per se*, os lingüistas circunscrevem-se aos limites de sua disciplina. Intencionalmente ou não, deixam transparecer certo descaso aos estudos computacionais da linguagem e resistem a cooperar com projetos de PLN e, sobretudo, não percebem a importância da sua contribuição para a construção do corpo de conhecimento interdisciplinar necessário para esse tipo particular de investigação. O lingüista Halvorsen (1989), mesmo ao defender a necessidade de interação entre Lingüística e PLN, avalia que a Teoria Lingüística, embora reúna dados significativos sobre o complexo competência-desempenho lingüístico humano, ao invés de incentivar a pesquisa do PLN, tem muitas vezes desestimulado seu desenvolvimento.

Finalmente, há que se observar que o problema de natureza terminológica e conceitual, embora mais acentuado nas relações multidisciplinares, ocorre também no interior da própria Teoria Lingüística. O emprego dos termos 'discurso' e 'texto', por exemplo, é um exemplo lapidar desse tipo de desencontro. Observe que, desta vez, as "confusões" localizam-se no âmbito dos estudos da

linguagem que ousaram investigar além dos limites da frase, confusões que têm gerado muitos debates e embates entre teóricos.[2]

2.2 *Desafio para os lingüistas*

Se os projetistas de PLN valem-se de justificativas como as apontadas anteriormente para se distanciarem dos lingüistas, estes também apresentam argumentos para não se engajarem em projetos de PLN.

Talvez o argumento mais difundido seja a constatação de que os engenheiros da linguagem ocupam-se com o desenvolvimento de sistemas de PLN muito rudimentares e desprovidos de qualquer fundamentação lingüística. Basta mencionar, por exemplo, os dicionários eletrônicos, os programas de exercícios sobre alguma disciplina acadêmica e as enciclopédias multimídia em CD-Rom, aplicativos cada vez mais popularizados. Nesses aplicativos, não há tecnologia lingüística alguma. Todos os objetos de natureza lingüística são simplesmente manipulados por técnicas de indexação e algoritmos que contornam os problemas computacionais postos pela complexidade das línguas naturais. Os "tradutores de bolso", por exemplo, não apresentam qualquer vestígio de PLN. Esses dispositivos, do tamanho de pequenas calculadoras ou canetas, são equipados com listas de palavras e expressões de línguas diversas, algumas frases e fragmentos de frases selecionados. O programa subjacente limita-se a manipular esses elementos: por meio de comparações, detecta as palavras equivalentes de línguas diferentes e, por meio de substituições, monta frases ou completa fragmentos de frases com as palavras e/ou expressões pré-armazenadas. Situação semelhante ocorre também com muitos corretores ortográficos, que se limitam a comparar palavra por palavra, sem executar qualquer análise morfológica ou gramatical.

Do ponto de vista teórico-metodológico, avaliar propostas e selecionar estratégias de trabalho no domínio do PLN transformam-se em problemas ainda mais complexos. Por ser um campo de estudos multidisciplinar, que vem sendo explorado por pesqui-

---

[2] Para uma apreciação das dificuldades e confusões causadas pela pluralidade de usos desses termos, remeto o leitor para Fávero e Koch (1983, p. 23) – para quem "as diferentes concepções de texto e discurso acabaram por criar uma confusão entre os dois termos, ora empregados como sinônimos, ora usados para designar entidades diferentes" e que atribuem parte das confusões à não existência, em algumas línguas, do termo 'discurso' –, Greimas e Courtés (1979). Há que se ressaltar que essas confusões acabam também por gerar uma série de denominações, não menos problemáticas, empregadas, por vezes, para demarcar fronteiras entre "lingüística frasal" e "lingüística transfrástica": *análise do discurso*, *lingüística textual*, *gramática* ou *sintaxe do texto*, *análise da conversação*.

sadores de áreas bastante diversas, depara-se com uma variedade de propostas, ferramentas e equipamentos computacionais, cuja sistematização torna-se praticamente impossível. Grande parte desses produtos de pesquisa, em geral assinados por não-lingüistas, encontra-se fragmentada e dispersa em incontáveis publicações e relatórios. São raros os pesquisadores que têm se preocupado com a apresentação sistematizada de um conjunto mínimo de conhecimentos já produzidos no campo.[3]

Nessa efervescência, encontram-se teorias e modelos em profusão: modelos estatísticos, modelos lógicos, teoria dos grafos, teoria dos conjuntos, teoria de modelos, teoria das linguagens formais, teoria dos algoritmos, teoria da complexidade, modelos de representação do conhecimento, entre outros. Parte desse aparato teórico-descritivo é também utilizada por lingüistas. Destaca-se, por exemplo, a teoria dos grafos e a teoria das linguagens formais, que, a partir da proposição da gramática gerativo-transformacional (Chomsky, 1957; 1965), são amplamente empregadas como esquemas de descrição e representação das regras e estruturas sintáticas das línguas. Há um lado positivo nessa diversidade. Do ponto de vista de recursos formais, imprescindíveis para o tratamento computacional das línguas naturais, há que se reconhecer que o conjunto desses trabalhos constitui um referencial rico, sugerindo modelos e sofisticadas técnicas de representação e manipulação de objetos lingüísticos.

O levantamento de projetos e de aplicativos de PLN revela, entretanto, um domínio de pesquisa difuso, controvertido e caótico; um domínio à espera de uma organização que assinale contornos mais claros e identifique seus objetos e crie condições mais favoráveis para que o estudo sistemático do PLN possa encontrar solo fértil para gerar projetos, de fato, integrados e interdisciplinares.

No âmbito da Teoria Lingüística, mesmo admitindo-se que não haja um referencial único, que modelos explícitos e completos ainda estejam para ser construídos e que lutas teóricas fazem parte de sua história (Leech, 1983), há que se concordar que existem parâmetros norteadores mínimos a que os lingüistas podem recorrer para ancorar suas investigações. Verifica-se um consenso mínimo em relação às características e funções fundamentais da linguagem humana – sua dupla-articulação, sua dependência estrutural, sua recursividade, suas funções representacional, expressiva, metalingüística, fática, intencional e textual – e em relação à terminologia e

[3] Obras clássicas como Grishman (1986), Allen (1987) e Gazdar e Mellish (1989) estão entre as poucas tentativas de reunir didaticamente os temas pertinentes ao PLN.

aos conceitos básicos – língua e linguagem, língua e fala, gramática, estrutura de constituintes, regras sintáticas recursivas, categorias sintáticas e funcionais, lexemas, categorias nucleares, papéis temáticos, esquema de subcategorização, restrições selecionais, casos morfológicos, categorias dêiticas e anafóricas, atos de fala, entre outros (Jakobson, 1977; Borba, 1984; Sells, 1985).

Essa base comum, resultante de uma longa tradição de estudos da linguagem, acaba por fornecer um universo de discurso comum, contendo uma metalinguagem e noções gerais, fato que se evidencia no discurso dos próprios teóricos, que constantemente recorrem a ela para construir suas análises. Uma leitura das propostas teóricas recentes é suficiente para constatar que termos que designam categorias e funções gramaticais, por exemplo, são tomados de empréstimo da gramática tradicional e, posteriormente, reciclados para rotular conceitos novos. Para exemplificar, destaco este trecho inicial da discussão de Marantz (1984, p. 1) sobre a natureza das relações gramaticais:

> Os lingüistas têm clareza sobre noções gerais, noções pré-teóricas, como 'antecedente de um pronome reflexivo', 'especificação de casos morfológicos', 'ordenação das palavras', 'agente de uma ação', e outros conceitos que parecem estar relacionados com a noção de 'sujeito'.

Envolver-se com o PLN, porém, implica estar disposto a compreender uma efervescência de teorias e técnicas, emprestadas das mais variadas áreas, e a também decifrar "formalismos algébricos muito mais bizarros".

Em se tratando de um campo de pesquisa em que parcelas de conhecimentos devem ser extraídas de domínios de estudos bastante diversos, o estudo do PLN nos coloca diante de uma outra dificuldade: uma metalinguagem fragmentada, um hermetismo terminológico e uma multiplicidade de termos exóticos. A terminologia que Shapiro (1990) discute para descrever os tipos de processamento ilustra a proliferação de termos, provenientes das Ciências da Computação e da Inteligência Artificial. Os três pares de modificadores *botton-up* (ascendente) / *top-down* (descendente), provenientes da sub-área "análise sintática", *forward* (para frente)/ *backward* (para trás), provenientes da sub-área "sistemas baseados em regras" e *data-driven* (direcionado para os dados) / goal-directed (direcionado para a meta), provenientes da sub-área "resolução de problemas", são utilizados para modificar os termos *chaining* (encadeamento), *inference* (inferência), *parsing* (análise gramatical), *processing* (processamento), *reasoning* (raciocínio) e

search (busca), gerando termos como: *botton-up/top-down parsing*, *forward/backward chaining*, *data-driven/goal-directed reasoning*, *data-driven/goal-directed processing*, *forward/backward search* e *forward/backward inference*.

É certo que a necessidade de se delimitarem novos conceitos e técnicas exije a proposição de novos termos, mais precisos e específicos. A sua proliferação assistemática, porém, acaba por criar dificuldades adicionais que precisam ser contornadas.

Lyons (1977), nos anos 70, já assinalava a necessidade de enfrentar tarefa semelhante no âmbito do estudo da semântica das línguas naturais. Ao buscar a construção de um referencial teórico comum para o estudo do significado e dos processos de comunicação lingüística, recorrendo aos trabalhos realizados no âmbito da Etnologia, Psicologia, Filosofia, Antropologia e Lingüística, conclui que a busca de sistematização de termos e conceitos transformou-se, para ele, em um pesadelo. Alertou, então, para o perigo de duas situações: (i) quando autores diversos atribuem acepções distintas a um mesmo termo e (ii) quando autores diversos empregam termos distintos para descrever fenômenos essencialmente idênticos. Após tentativas frustradas de encontrar uma maneira de compatibilizar as discrepâncias, Lyons optou por simplesmente apontá-las e, na medida do possível, selecionar termos que lhe pareciam mais adequados aos seus propósitos.

Além disso, termos como "conhecimento", "inferência", "inteligência", "raciocínio", "pensamento", "capacidade", "compreensão", "interpretação", "significado", entre outros, próprios do universo humano, no âmbito do PLN, são freqüentemente transportados para o universo dos computadores. Nesse universo, as máquinas pensam, são consultoras, estabelecem diagnósticos, dão consultoria e conselhos, conversam não só entre si mas com os usuários humanos, falam línguas, ensinam, planejam tarefas e até aprendem.

Minsky (1968, p. 2) até mesmo chega a defender essa transposição:

> Alguns leitores podem se sentir incomodados por eu deliberadamente usar termos do universo psicológico, tais como "significado", que, comumente, não são empregados na descrição do comportamento de máquinas. Mas minha opinião é de que o uso desses termos mentalistas não é uma simples analogia. O fato é que os programas de computador, aqui descritos, confirmam a validade e fertilidade da revolução intelectual que resultou da descoberta de que, pelo menos, algumas descrições mentalistas dos processos de pensamento podem ser transformadas em especificações para a construção de máquinas.

A antropomorfização da máquina pode ser justificável do ponto de vista da inteligibilidade das explicações. Essa estratégia discursiva, no entanto, acaba gera também interpretações que contribuem para formar a imagem de que pesquisar PLN é adotar uma visão mecanicista do homem, ou ainda, que o PLN é, por definição, um campo desprovido de conceitos e termos precisos para descrever seus objetos.

A precária troca de trabalhos sobre o PLN entre os lingüistas e os projetistas de sistemas de PLN fica também evidente nas raras publicações traduzidas para o português. Os textos, que em sua massiva maioria estão publicados em inglês, quando são traduzidos, além de trazerem termos bizarros, apresentam imprecisões e confusões conceituais.

Merece destaque a denominação domínio "processamento automático das línguas naturais", expressão que venho empregando como equivalente à expressão inglesa "*automatic natural language processing*", cunhada pelos estudiosos do campo da Inteligência Artificial e das Ciências da Computação. Essa expressão exige esclarecimento, uma vez que a denominação corrente no português é "processamento de linguagem natural". Além disso, observe-se que termo inglês "*natural language processing*", embora seja também empregado para nomear o PLN, é uma expressão ambígua, pois nomeia também o estudo do processamento da linguagem humana, no âmbito da Ciência Cognitiva (Garret, 1990).

O termo "processamento automático" parece não ser motivo de controvérsias. No sentido usual, refere-se à utilização de computadores para a estruturação e manipulação de símbolos sem a intervenção humana. Essas operações são executadas segundo representações precisas e explícitas, implementadas por meio de programas escritos em uma linguagem de programação apropriada.

Já a adoção do termo "linguagem", exige melhor delimitação, porque os pesquisadores da área das Ciências Exatas, acostumados a trabalhar com linguagens formais, acabam por empregá-lo com sentido distinto daquele fixado pela tradição lingüística, gerando confusões desnecessárias e, principalmente, revelando desconhecimento de conceitos lingüísticos de fundo. Exemplifico o problema citando alguns segmentos da tradução para o português de um livro clássico sobre inteligência artificial. No capítulo "Compreensão de (*sic*) Linguagem Natural" (Rich, 1983, p. 344-406), observam-se quatro ocorrências do termo "linguagem", que está sendo equivocadamente empregado no lugar do termo "língua":

> A capacidade de se comunicar em um tipo de linguagem natural, seja ela inglês ou tagalog, parece ser considerada, às vezes, a aspiração máxima da raça humana […] Os mapeamentos muitos-a-um são mais comuns, particularmente quando estiverem mapeando de uma linguagem natural… para uma pequena representação-alvo simples […] Por outro lado, em muitas frases de linguagem natural, a mudança de uma única palavra pode alterar não apenas um único nó da interpretação, mas toda sua estrutura […] Para fazer a análise sintática de uma frase, é necessário utilizar uma gramática que escreva a estrutura de cadeias de uma linguagem em particular.

Nesse contexto, fica evidente que a autora refere-se a uma língua em particular e não à linguagem em geral. Lyons (1981, p. 16) esclarece essa questão:

> Diversas línguas européias têm duas traduções, e não uma, para o vocábulo inglês *language*: haja vista o francês *langage*: *langue*, o italiano *linguaggio*: *língua* e o espanhol *lenguaje*: *lengua*. Em cada um dos casos, a diferença entre as duas palavras está correlacionada, até certo ponto, com a diferença entre os dois sentidos da palavra inglesa *language* […] o inglês permite a seus falantes dizer de alguma pessoa que não só '*he possesses a language*' ['ele possui uma língua'] (inglês, chinês, malaio, suaíli, etc.), mas que '*he possesses language*' ['ele é dotado de linguagem'].[4]

O termo 'linguagem artificial' aplica-se, talvez metaforicamente, aos sistemas de notação ou cálculo elaborados por matemáticos, cientistas da computação e lógicos para fins específicos, que incluem, por exemplo, as linguagens de programação de computadores e a linguagem da lógica simbólica, que apropriadamente recebem o nome de linguagens artificiais.[5]

Grande parte da metalinguagem dos trabalhos de PLN, como uma colcha de retalhos, contém fragmentos de outras metalinguagens, constituindo um obstáculo adicional. A interpretação e a utilização dos termos e conceitos exigem cuidado especial para que não se mergulhe em um caos terminológico-conceitual.

---

4 Akmajian et al. (1986, p.6) comentam a estranheza manifestada por alunos americanos quando descobrem que o termo inglês "*language*" possui também o sentido genérico de "faculdade da linguagem". Acrescentam que, para os americanos, a distinção entre língua e linguagem parece ser praticamente desconhecida fora do círculo dos lingüistas, o que vem reforçar o cuidado com a precisão terminológico-conceitual. No português, o problema do emprego dos termos 'língua' e 'linguagem' nem mesmo se coloca, uma vez que cada um deles reveste conceitos distintos.

5 O termo 'linguagem', por ser de aplicação mais geral que o termo 'língua', é licitamente usado para denotar os sistemas de comunicação em geral, naturais e artificiais, entre seres humanos ou não: as linguagens de programação, a linguagem das abelhas, a linguagem corporal humana, a linguagem do trânsito, etc.

Nem sempre, porém, é tarefa fácil encontrar um termo do português que corresponda, com exatidão, ao termo criado em inglês para fazer referência aos novos conceitos. O segmento de texto, a seguir, traduzido da mesma obra mencionada acima, oferece uma ilustração dessa dificuldade (Rich, 1983, p. 347):

> Há três fatores principais que contribuem para a dificuldade de um problema de compreensão: [a] A complexidade da representação-alvo em que o casamento estiver sendo feito; [b] O tipo de mapeamento: um-a-um, muitos-a-um, um-a-muitos ou muitos-a-muitos; [c] O nível de interação dos componentes da representação de origem.

Os termos "casamento" e "mapeamento" são, nesse trecho, as traduções propostas para as expressões inglesas "*matching*" e "*mapping*", respectivamente. Eles, entretanto, não refletem os conceitos do domínio de que foram extraídos. O termo "*mapping*" é empregado para denotar uma função, transformação, projeção ou correspondência entre duas ou mais estruturas (Partee et al., 1993). Já o termo "*matching*" refere-se a um dos processos mais estudados no âmbito da inteligência artificial: o "*pattern matching*" (Slagle & Gini, 1990). Os termos "configuração", "projeção" e "unificação" parecem ser traduções mais apropriadas para os termos "pattern", "mapping" e "matching", respectivamente.

2.3 *Desafio para ambos*

Além dessas razões técnicas, grande parte dos ataques recíprocos entre os pesquisadores parece ser decorrência de um outro fato, também amplamente percebido em nossos meios acadêmicos: a imagem estereotipada e distorcida que os pesquisadores formam uns dos outros, sobretudo, se trabalham em domínios de conhecimento diversos.[6] Não é difícil constatar que uma simples conversa entre colegas de áreas distintas é, não raro, pontuada por esses julgamentos pré-concebidos.

Lyons (1979), no *Prefácio* de sua obra clássica *Introdução à Lingüística Teórica*, revela os dois estereótipos que há muito têm sido atribuídos a pesquisadores das duas áreas. Num momento em que os modelos formais de gramática passavam a ser o centro das investigações lingüísticas, Lyons advertia os leitores, em especial aqueles cuja formação intelectual se apoiava mais nas Humanidades, para o fato de que eles deveriam estar preparados para fazer

---

[6] Esse fato é reflexo do modelo compartimentado de pesquisa que ainda predomina em nossos meios acadêmicos, construindo barreiras entre as áreas do conhecimento e distanciando os pesquisadores.

"um certo esforço intelectual com respeito ao uso de símbolos e de fórmulas".[7] Além do rótulo explícito, "colegas dos números", é possível ler nas entrelinhas do *Prefácio* que os pesquisadores das "ciências exatas" são caracterizados como indivíduos pouco intuitivos no que se refere à "apreciação dos vários matizes da língua", academicamente mal formados para o estudo das línguas, uma vez que parecem desconhecer os fundamentos históricos e filosóficos dos estudos da linguagem, e responsáveis pela criação e proliferação de uma quantidade de "símbolos e fórmulas arcanas", cuja compreensão exige alta capacidade cognitiva. Já os pesquisadores das "humanidades" são caracterizados como aqueles de "hábitos mentais mais voltados para as letras", mais capazes de fazer uma "apreciação intuitiva dos vários matizes da língua" e conhecedores dos fundamentos históricos e filosóficos dos estudos da linguagem, porém, pouco acostumados a lidar com formalismos.

O uso dos rótulos "letras" e "números" e as associações implícitas "humanidades-pesquisa não científica" e "ciências-pesquisa científica", além de marcarem os pesquisadores de modo estereotipado e ilustrarem com precisão a divisão em compartimentos da academia, nitidamente cristalizada, revelam também que, por questão de poder, muitos pesquisadores negligenciam as questões diretamente relacionadas à compreensão do seu próprio objeto de estudo devido a lutas teóricas, quer no interior de um mesmo domínio do conhecimento quer no confronto de domínios diferentes.

Assim, a tímida interação entre os engenheiros e os cientistas da linguagem, alimentada por desencontros e descrédito mútuo, evidencia (i) o desconhecimento que o pesquisador de uma área demonstra ter do trabalho desenvolvido na outra, (ii) as tentativas frustradas de compreensão das metalinguagens, dos conceitos, dos métodos e das técnicas específicos de cada domínio, (iii) a disseminação de imagens estereotipadas, (iv) a existência das lutas teóricas e (vii) o descaso com a interdisciplinaridade.

Enquanto lingüistas lutam para introduzir sofisticados recursos da informática em suas pesquisas lingüísticas, projetistas de PLN lutam para formatar as línguas naturais em seus computadores sem uma fundamentação lingüística adequada.

---

[7] Hoje, passados mais de 30 anos, verifica-se que advertência semelhante precisa ser feita ao se abordar os estudos sobre o PLN. Desta vez, a advertência não é direcionada apenas àqueles de "formação humanística", mas aos próprios lingüistas. Mesmo estando familiarizados com os múltiplos formalismos, que hoje é um lugar-comum em qualquer teoria lingüística, os lingüistas precisam estar preparados para conseguir decifrar um volume considerável de representações, formalismos e o próprio jargão (objetos muito mais "arcanos") que fazem parte do universo do PLN.

Essa duplicação desnecessária de esforços não só dificulta a descoberta de soluções, que seguramente seria agilizada com o incentivo de trabalho solidário. A desvinculação das duas áreas é preocupante, porque contribui para aumentar os desencontros e minimizar a importância do papel dos lingüistas na proposição e no desenvolvimento de projetos de PLN que, em sua grande maioria, resultam de iniciativas de não-lingüistas e acabam por ficar circunscritas a instituições ou departamentos que investem na pesquisa tecnológica.

3 Cooperar é preciso

Diante desse quadro de desencontros, fica evidente que não é tarefa nada trivial a construção de um universo de discurso comum que possibilite o diálogo entre os cientistas os engenheiros linguagem e, a partir desse universo, a criação de estratégias de trabalho interdisciplinar e solidário que viabilizem a criação de núcleos de pesquisa com esse perfil. Como Starosta (1991, p. 195), acredito, entretanto, na possibilidade de se estabelecer uma cooperação colaborativa entre ciência e engenharia da linguagem e que essa produza teorias lingüísticas mais realista e aplicações computacionais mais robustas e precisas. Com essa parceria, os estudos do PLN e os estudos lingüísticos poderão beneficiar-se com o trabalho interdisciplinar, que incentive o envolvimento dos vários especialistas na busca de soluções integradas. Pesquisas dessa natureza poderão ainda contribuir para a proposição de modelos lingüísticos mais completos, explícitos e operacionais e, conseqüentemente, mais apropriados para receberem o tratamento computacional.

O desafio é, portanto, viabilizar a formação de um domínio de pesquisa interdisciplinar com potencial tecnológico, integrando recursos teóricos e técnicas de investigação desenvolvidos no âmbito de um conjunto de disciplinas matrizes entre as quais a Teoria Lingüística deve desempenhar papel fundamental.

Rich (1985) demonstra que tanto a Inteligência Artificial como as Humanidades podem beneficiar-se com o estudo do PLN, uma vez que o problema de emular/simular o complexo conhecimento-desempenho lingüístico humano em uma máquina exige que o estudioso a explicite com clareza e precisão os elementos e princípios constitutivos das línguas naturais.

Lyons (1979) combate a visão separatista de pesquisa entre ciências e humanidades. Para ele, são poucos os ramos do conhecimento que sofrem mais do que a Lingüística com a separação entre as duas áreas, dicotomia que ainda se mantém nos currículos da maioria das nossas escolas e universidades. Para ele, a Teoria Lin-

güística deve aproveitar-se eqüitativamente das abordagens "tradicional" e "científica" da língua, considerando este resultado da intersecção entre os estudos da linguagem e os estudos desenvolvidos no âmbito da Lógica Formal, da Análise Computacional e da Teoria dos Autômatos.

Notável é constatar que projetistas de PLN como Winograd (1972, p. 41) também preocupam-se com o embasamento lingüístico do estudo do PLN:

> Quando todas as tentativas para salvar o empreendimento da tradução automática falharam, ficou patente que foi muito prematuro, por parte dos pesquisadores, tentar abarcar toda a língua inglesa sem buscar fundamentação mais sólida na teoria lingüística e sem compreender as propriedades matemáticas das gramáticas.

3.1 *A concepção lingüisticamente motivada de PLN*

Para esse precursor do estudo do PLN,[8] os elementos essenciais para o estudo do PLN e, sobretudo, para a indispensável ancoragem lingüística resumem-se no que segue:

> Assumimos que um computador não poderá satisfatoriamente emular uma língua natural se não conseguir 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 precisa ter um modelo simples de sua própria mentalidade. Ele precisa ser capaz de lembrar seus planos e ações, discuti-los [com o usuário] e executá-los. Ele deve participar de um diálogo, respondendo, com ações e frases, às frases digitadas em inglês pelo usuário; [ele deve] solicitar esclarecimentos quando seus programas heurísticos não conseguem compreender uma frase com a ajuda das informações sintáticas, semânticas, contextuais e do conhecimento de mundo físico previamente representadas no sistema.

Além de evidenciar o complexo de conhecimentos e habilidades envolvidos no processo de comunicação verbal, e que precisam estar representados dentro de um sistema de PLN, Winograd (1972, p. ix) nos ensina que pesquisar o PLN pode ser também um modo de investigação científico dos próprios fatos da língua:

---

8 A importância de Winograd para o desenvolvimento do PLN decorre de sua tese de doutorado, defendida em 1970, no MIT. Nesse trabalho, Winograd criou um sistema computacional que passou a ser o marco do estudo científico do PLN. Com esse sistema, denominado SHRDLU, também conhecido como "o mundo dos blocos", conseguia mostrar para a comunidade científica que a interação homem-máquina era realmente possível por meio de uma língua natural. Sob a forma de uma representação gráfica no monitor do computador, o sistema simula o braço de um robô que manipula um conjunto de blocos de formas e cores diferentes sobre a superfície de uma mesa conforme as instruções digitadas em inglês no teclado do computador.

> Todo mundo é capaz de compreender uma língua. A maior parte do tempo de nossas vidas é preenchida por atos de fala, leituras ou pensamentos, sem sequer notarmos a grande complexidade da linguagem. Ainda não sabemos como 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 por ela codificados.

3.2 *A evolução do estudo do PLN*

Nesse meio século de pesquisas, é possível fazer o seguinte quadro-síntese da evolução do estudo do PLN em termos do grau de sofisticação lingüística alcançado.

Quadro 1. Evolução do estudo do PLN.

Década	Foco da Investigação	Conquistas
50	Explorações: tradução automática	sistematização computacional das classes de palavras descritas nos manuais de gramática tradicional; identificação computacional de constituintes oracionais.
60	Formalizações: novas aplicações e criação de formalismos	primeiros tratamentos computacionais das gramáticas livres de contexto; criação dos primeiros analisadores sintáticos; primeiras formalizações do significado em termos de *redes semânticas*.
70	Criação do nicho de pesquisa: consolidação do PLN	implementação de parcelas das primeiras gramáticas e analisadores sintáticos baseados na *gramática gerativo-transformacional*; busca de formalização de fatores pragmáticos e discursivos.
80	Busca da precisão: sofisticação dos sistemas	desenvolvimento de teorias lingüísticas motivadas pelos estudos do PLN como, por exemplo, a *gramática sintagmática generalizada* e a *gramática léxico-funcional*.
90	Busca da precisão e robustez: sistemas baseados em representações do conhecimento no tratamento estatístico de massa de textos	desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos lingüísticos e extralingüísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos lingüísticos para os quais os sistemas são projetados. ressurgimento da lingüística de corpus e do tratamento estatístico de entidades e processos lingüísticos.

Nesse domínio, pesquisas interdisciplinares encontram solo fértil para germinar. De fato, Sanders e Sanders (1989, p. 30), também evidenciando as dificuldades de interlocução existentes entre pesquisadores de áreas distintas, reconhecem a importância do trabalho solidário:

> Os cientistas da computação sabem propor e gerenciar projetos de software. Eles dispõem de equipamentos e ferramentas de programação de vanguarda; as linguagens simbólicas são seu material de trabalho. Por outro lado, entretanto, eles freqüentemente não dominam os conhecimentos lingüísticos [...] É evidente que trabalho de equipe é condição essencial. Entretanto, a comunicação entre especialistas diversos não é tarefa fácil [...]

Assim, a busca de estratégias de trabalho que possibilitem a aproximação dos diversos especialistas, a produção efetiva de conhecimento interdisciplinar e a aplicação desse conhecimento no estudo do PLN são condições essenciais para a sua solidificação.

3.3 *Metodologia de pesquisa do PLN*

A esta altura das discussões, torna-se evidente a necessidade de se somarem competências específicas para a realização do empreendimento. A grande questão que se coloca é como criar uma estratégia de pesquisa integrada e um sistema computacional apropriados para o estudo do PLN.

A estratégia de pesquisa do PLN que proponho envolve equacionar os problemas em três **Domínios** de investigação: **Lingüístico**, **Lingüístico-Computacional** e **Computacional**. A proposta é motivada pelo paradigma de desenvolvimento de sistemas computacionais que visam à construção de bases de conhecimento (Hayes-Roth, 1990) e na programação automática (Biermann, 1990). Esses sistemas são projetados para aplicar automaticamente complexos de conhecimentos neles armazenados no processo de resolução de problemas (Amarel, 1990).

A metodologia de construção de sistemas especializados dessa natureza pressupõe a especificação dos tipos de conhecimentos que os especialistas emulados no sistema possuem, como esse conhecimento é adquirido, armazenado, acessado e aplicado (Schank & Riesbeck, 1981, p. 2). Assim, investigar como um sistema de computador pode emular parcelas da competência e da atuação de um sintaticista, por exemplo, pressupõe especificar os conhecimentos e habilidades que esse especialista possui. De modo análogo, estudar como um sistema de computador pode

emular parcelas da competência e do desempenho lingüísticos humanos pressupõe especificar de modo sistemático e formal os conhecimentos e as habilidades que os falantes, especialistas nesse domínio, possuem.

Assumindo a concepção de PLN de Winograd, conclui-se que emular aspectos de língua natural pressupõe equipar um sistema de PLN com vários sistemas de conhecimento e fazê-lo emular uma série de atividades cognitivas:

- possuir um "modelo simples de sua própria mentalidade";
- possuir um "modelo detalhado do domínio específico do discurso";
- possuir um modelo que represente "informações morfológicas, sintáticas, semânticas, contextuais e do conhecimento de mundo físico";
- "compreender o assunto que está em discussão";
- "lembrar, discutir, executar seus planos e ações";
- participar de um diálogo, respondendo, com ações e frases, às frases digitadas pelo usuário;
- solicitar esclarecimentos quando seus programas heurísticos não conseguirem compreender uma frase.

Observe-se que a analogia que estou construindo entre investigar o PLN e investigar sistemas especialistas permite conceber a primeira atividade como a atividade de construir um tipo particular de sistema de processamento automático de informações de natureza lingüística e de informações lingüisticamente relevantes, cuja especialidade inclui realizar tarefas como revisão ortográfica de textos, análise gramatical, tradução de frases ou textos, interação verbal com o usuário e auxílio especializado a pesquisadores na própria construção de modelos lingüísticos. Nesse sentido, o estudo do PLN é concebido como um tipo de "engenharia do conhecimento lingüístico" e beneficia-se da metodologia da "engenharia do conhecimento" (Hayes-Roth, 1990).

De modo semelhante ao processo de investigação de um sistema de conhecimento (do inglês *knowledge system*), como o processo de mineração, o estudo do PLN compreende a realização de, no mínimo, três etapas investigação: a "extração do solo" (isto é, a explicitação dos conhecimentos e habilidades lingüísticas), a "lapidação" (isto é, a representação formal desses conhecimentos e habilidades) e a "incrustação" (isto é, a construção do programa de computador que codifica essa representação). O Quadro 2 sintetiza essas etapas e especifica os resultados esperados em cada uma delas.

Quadro 2. Etapas de construção de um sistema de PLN.

Tarefas	Resultados
Explicitação do conhecimento	Descrições lingüísticas precisas
Representação formal do conhecimento	Representações lingüístico-computacionais
Codificação computacional do conhecimento	Implementação computacional

Os estudos da linguagem desenvolvidos no âmbito da Ciência Cognitiva também propõem três níveis de abordagem do processamento humano da linguagem que correspondem às três fases acima (Lasnik, 1990: xvii-iii):

> A idéia central da ciência cognitiva moderna é que o sistema cognitivo humano pode ser entendido como um computador gigante que executa cálculos complexos. [...] No caso da linguagem humana, por exemplo, o nível de implementação corresponde à análise neurológica das estruturas e conexões do cérebro que estão subjacentes ao uso das línguas. O nível de representação e algoritmos focaliza o processamento da informação pelo sistema e o formato do conhecimento lingüístico armazenado na memória. [...] No nível computacional, a língua é analisada em termos gramaticais e suas propriedades estruturais são expostas. [...] É fundamental compreender que [...] os três diferentes níveis de análise estão ligados, isto é, os fatos e os princípios descobertos em um nível contribuem para análises nos outros níveis. Por exemplo, o conhecimento da gramática de uma língua (descrito no nível 2) nos dá pistas sobre o tipo de algoritmo necessário para interpretar e produzir frases.

Por fim, Barton, Berwick e Ristad (1987, p. 96-7) esclarecem que a teoria da competência lingüística, inclui-se no "nível computacional" e deve explicar quais são as estruturas calculadas e por quê, ignorando as limitações de memória, as mudanças de atenção ou interesse, e os erros. Assim, somando as questões programáticas sobre os estudos da linguagem colocadas por Chomsky (1986, p.3) e as considerações sobre a "competência pragmática", propomos a Figura 1, que resume o equacionamento do estudo de PLN.

Figura 1. Correlação entre os Domínios de investigação do PLN, os problemas centrais de cada domínio e os recursos teórico-metodológicos essenciais, mas não exaustivos, para resolvê-los.

DOMÍNIOS	PROBLEMAS	RECURSOS [9]
Lingüístico ↓↑ Lingüístico-Computacional ↓↑ Computacional	Descrever o conhecimento e o uso lingüísticos ↓↑ Representar os conhecimentos do domínio anterior ↓↑ Codificar em uma linguagem de programação as representações propostas no domínio anterior	Teorias Lingüísticas da Competência e do Desempenho ↓↑ Linguagens Formais de Representação ↓↑ Linguagens de Programação e Sistemas de Computadores

Como ressalta Halvorsen (1989, p. 201), o estudo do PLN tem, de fato, procurado:

> construir a ponte entre a teoria da competência e o tipo de desempenho lingüístico atribuído às máquinas, transformando a teoria lingüística em algoritmos que, ao mesmo tempo, simulam o comportamento lingüístico e obedecem às restrições e generalizações previstas pela teoria lingüística e pelas gramáticas [das línguas particulares].

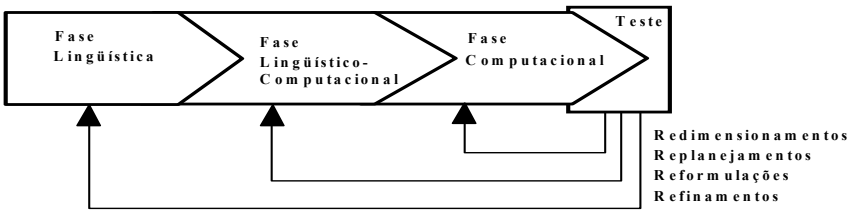
A explicitação do conhecimento e do uso lingüísticos envolve questões do Domínio Lingüístico, uma vez que é nele que os fatos de língua e de uso são especificados. Conceitos, termos, regras, princípios, estratégias de resolução de problemas e formalismos lingüísticos são os elementos trabalhados. No Domínio Lingüístico-Computacional, questões referentes à escolha ou à proposição de sistemas de representação, que incluem, por exemplo, o cálculo proposicional e de predicados, as redes semânticas, as regras de reescrita e os *frames*, bem como as estratégias de codificação dos elementos trabalhados no domínio anterior entram em foco. No Domínio Computacional, além das questões que envolvem a implementação das representações por meio de programas de computador, há questões que dizem respeito à montagem do próprio sistema computacional em que os programas são alojados.

[9] Os principais recursos disponíveis para o estudo do PLN estão resumidos na Figura 3.

Em suma, a investigação nos três Domínios, por sua vez, pode ser reinterpretada como compreendendo três fases solidárias e sucessivas do desenvolvimento de um sistema PLN específico. Na Fase Lingüística, estabelece-se a construção do corpo de conhecimentos sobre a própria linguagem, dissecando e compreendendo os fenômenos lingüísticos necessários para o desenvolvimento do sistema. Nesta fase, a análise dos fenômenos lingüísticos é elaborada em termos de modelos e formalismos desenvolvidos no âmbito da Teoria Lingüística. Na Fase Lingüístico-Computacional procede-se à construção conceitual do sistema, que envolve a seleção ou proposição de sistemas formais de representação para os resultados alcançados na fase anterior. Nesta fase, projetam-se as representações lingüísticas e extralingüísticas em sistemas formais computacionalmente tratáveis. Na Fase Computacional, planeja-se o sistema e codificam-se as representações elaboradas na fase anterior em linguagens de programação. Nesta fase, além das representações da fase anterior serem codificadas em programas computacionais, estudam-se a integração conceitual e física dos vários componentes do sistema, bem como o desenho e a implementação do ambiente computacional em que o sistema será desenvolvido.

Conforme ilustra a Figura 2, as três fases devem ser desenvolvidas sucessiva, progressiva e ciclicamente: as representações parciais resultantes das duas primeiras fases podem ser implementadas e, finalmente, testadas, completando, assim, um ciclo.[10] Dessa forma, testes de adequação e desempenho são realizados para aferir o aprimoramento dos resultados alcançados em cada fase.

Figura 2. Fases de construção de um sistema de PLN.



[10] Essa decomposição modular de um problema complexo em subproblemas espelha também uma estratégia de trabalho bastante difundida e profíqua nos estudos computacionais (RICH, 1983). Bresnan (1981; 1982) emprega estratégia semelhante na construção da Teoria Léxico-Funcional.

Assim, projetar um sistema de PLN envolve essencialmente (i) especificar, (ii) representar e (iii) codificar sistematicamente informações (lingüísticas e extralingüísticas), mecanismos de inferência e de controle dessas inferências, e, finalmente, (iv) projetar o sistema computacional (incluindo *software* e *hardware*) para o desenvolvimento e teste do próprio sistema. Isso equivale a dizer que é preciso construir a representação de um complexo "competência-desempenho lingüístico e metalingüístico artificial" e transformar tudo isso em um programa de computador

Quadro 3. Exemplo do tratamento lingüístico-computacional de uma configuração sintática.

FASE LINGÜÍSTICA	**DESCRIÇÃO INFORMAL** "Uma frase pode ser composta pela concatenação de um Sintagma Nominal com um Sintagma Verbal. O Sintagma Nominal é o Sujeito da frase. O Sintagma Verbal é o predicado da frase. O Sujeito e o Verbo têm os mesmos traços de Número e Pessoa. O Caso Gramatical do Sujeito é o Nominativo e o Verbo encontra-se na Forma Finita."
	FORMALISMO LINGÜÍSTICO F → SN SV (↑ SUJEITO) = ↓ ↑ = ↓
FASE LINGÜÍSTICO-COMPUTACIONAL	**REPRESENTAÇÃO COMPUTACIONAL** **Regra Sintática:** **F → SN SV** **Especificações:** **<SN pessoa> = <SV pessoa>** **<SN número> = <SV número>** **<SN caso> = nominativo** **<SV forma verbal> = finita**
FASE COMPUTACIONAL	**IMPLEMENTAÇÃO NA LINGUAGEM PROLOG**[11] f(P0,P):-sn(Pessoa,Número,Caso, P0,P1), sv(Pessoa,Número,Caso,P1,P).

O exemplo ilustrativo, esquematizado no Quadro 3, auxilia compreensão da proposta. Nesse esquema, partimos da descrição informal de um tipo específico de configuração sintática oracional e chegamos a um programa de computador que a codifica.

[11] Cf. Sterling & Shapiro, 1986; Pereira & Shieber, 1987; Clocksin & Mellish, 1987.

4 A essência interdisciplinar do PLN

Desde a Antigüidade, textos e mais textos vêm registrando um corpo de conhecimentos sobre os fenômenos lingüísticos das mais diversas perspectivas, refletindo idéias, preocupações e visões de mundo específicas de cada época. Em cada momento, as lentes dos observadores estão direcionadas para determinados aspectos dos objetos lingüísticos, ocultando outros, visíveis sob outras óticas.

Os lingüistas, num certo momento, ao observarem os fenômenos da linguagem com as lentes do gerativismo, corrente de pesquisa gramatical dominante na segunda metade do século XX (Harlow & Vincent, 1989), focalizaram suas lentes até os limites da frase. Qualquer fenômeno a ela transcendente fugia desse alcance. A fonologia e a morfologia foram minimizadas e a sintaxe ocupou o centro das discussões. A semântica, por sua vez, continuo posta de lado. Afinal, como salienta Lyons (1979, p. 425-6), para a Lingüística Moderna, o estudo do problema do significado era tarefa para psicólogos, filósofos, lógicos, antropólogos e sociólogos.

A partir da década de 70, com outras lentes, algumas emprestadas de estudiosos de outras disciplinas, o espaço da semântica, bem com tantos outros espaços de investigação foram demarcados: a *análise do discurso* (Prince, 1988), a *pragmática* (Levinson, 1983; Leech, 1983; Horn, 1988), os *postulados conversacionais* (Kempson, 1988), os *atos de fala* (Sadok, 1988), entre outros.

Fatos como esses permitem reiterar o que já afirmara em Dias-da-Silva (1990): a colocação de problemas, a seleção de questões e a busca de soluções não são determinadas exclusivamente pela natureza do objeto sob investigação. Cada tipo de abordagem, com seus métodos próprios, além de definir determinados contornos do objeto, acabam também por propiciar o nascimento de novos domínios de estudo. A caracterização de novos objetos ou de lentes novas é, na maioria das vezes, fruto de influências de outras áreas do saber sobre a Lingüística.

Um exemplo significativo dessas contribuições recíprocas é encontrado em Chomsky (1957). Recorrendo à Teoria dos Autômatos (Korfhage, 1966), desenvolvida por matemáticos, Chomsky propôs o modelo formal de análise gramatical que revolucionou os estudos da linguagem. Como observa Lyons (1976, p. 63):

> No que diz respeito à lingüística, o passo revolucionário dado por Chomsky foi o de recorrer a esse ramo da matemática [teoria dos autômatos finitos e teoria das funções recursivas], aplicando-o às línguas naturais, como o inglês, e não a línguas artificiais, construídas por lógicos e por cientistas especializados em computação.

Acrescente-se a esse avanço a sua ressonância nos estudos matemáticos. Ao estudar a possibilidade de criar modelos formais de gramática para descrever a sintaxe das línguas naturais, Chomsky criou também uma nova área de investigação, essencial para os estudos computacionais: o estudo das *linguagens formais* (Sudkamp, 1991), que, por sua vez, foi decisivo para a criação das linguagens de programação, dos compiladores e dos interpretadores. Como conseqüência, serviu também de estímulo para o estudo do PLN, que até a década de 70, ainda sofria o efeito negativo das experiências mal-sucedidas com a tradução automática (Nirenburg et al., 1992).

Assim, o estudo das linguagens formais forneceu o contexto para o nascimento de uma nova área da Ciência da Computação, área que ficou conhecida como Lingüística Computacional, e que, apesar do nome, não deve ser considerada um desdobramento da Lingüística. Sua lente, até meados da década de 60, centrou-se exclusivamente no estudo das linguagens formais e das linguagens de programação (Ballard & Jones, 1990, p. 133). Mesmo hoje, com o amadurecimento crescente dos estudos sobre o PLN, a Lingüística Computacional concentra-se em um aspectos específicos do processamento lingüístico: *estudo de algoritmos para análises morfológica* e *gramatical* (Earley, 1970; Kay, 1985; Hearn et al., 1980, Klavans, 1989). O estudo dos sistemas de representação e dos procedimentos computacionais de interpretação semântica e pragmático-discursiva, em geral, ultrapassam o seu domínio (Grishman, 1986; Nirenburg et al., 1992).

Com o estímulo proveniente da Lingüística e com influências diretas da Filosofia da Linguagem e da Psicologia, o estudo do PLN passou a abordar os mesmos temas dessas disciplinas matrizes: morfologia, sintaxe, semântica, pragmática, discurso, texto, aquisição da linguagem, entre outros (Carbonell & Hayes, 1990). Exemplos significativos que atestam as contribuições recíprocas que passaram a existir entre este e os estudos lingüísticos são a proposição de modelos diversos: modelos de processamento automático de estruturas gramaticais (Frazier & Fodor, 1978; Fodor & Frazier, 1980); modelos de rede de transição ampliada projetados para representar o processo de análise sintática valendo-se da gramática gerativo-transformacional (Woods, 1970); modelos de codificação e implementação de parcelas da gramática funcional proposta por Halliday (Winograd, 1972);[12] modelos computacionais dos atos de fala (Cohen & Perrault, 1979; Allen & Perrault,

[12] Cf. Halliday & Hasan (1976) e Halliday (1985).

1980); analisadores gramaticais fundamentados na gramática gerativo-transformacional (Marcus, 1980); modelos de gramática computacionalmente motivados (Bresnan, 1982; Dik, 1997); modelos matemáticos das línguas naturais (Perrault; 1984); modelos computacionais de geração de textos (McKeown, 1985; Appelt, 1985); modelos formais de interpretação semântica (Dowty et al., 1981; 1985); analisadores gramaticais fundamentados na teoria chomskiana dos princípios e parâmetros (Pritchett, 1988); modelos computacionais da teoria da referência (Kronfeld, 1990); modelos computacionais de interpretação semântica (Hirst, 1992).

Assim como a Lingüística, a Inteligência Artificial também enriqueceu o estudo do PLN e, por meio deste, avançou seus próprios conhecimentos. Entre os temas mais importantes nas discussões sobre a criação de sistemas capazes de processar as línguas naturais estão: as *estratégias de resolução de problemas* (Amarel, 1990), as técnicas de *representação do conhecimento* (Brachman & Levesque, 1985; Minsky, 1975) e as teorias que estudam sofisticados *sistemas de inferência* (Reyter, 1987; Hobbs et al., 1990; Carpenter & Thomason, 1990); o modelo de *redes semânticas*, criado para a representação da estrutura conceitual que serve de ancoragem para a estruturação do léxico (Quillian, 1968; Jackendoff, 1990; Pustejovsky & Boguraev, 1991; Pustejovsky, 1995; Miller & Fellbaum, 1991) e as técnicas empregadas pela *engenharia do conhecimento* (Hayes-Roth, 1990).

Mesmo que, historicamente, a Inteligência Artificial e a Lingüística Computacional,[13] ambas consideradas ramificações da Ciência da Computação (Ballard & Jones, 1990; Nirenburg et al., 1990), tenham tomado para si o estudo do PLN, a concepção de Winograd e os trabalhos mencionados o colocam como um empreendimento interdisciplinar. Dessa nova perspectiva, o PLN não se constitui em objeto específico desta ou daquela área do conhecimento, mas sim um objeto complexo e multifacetado, cuja compreensão tem se revelado potencialmente promissora e responsável por significativas influências recíprocas.

---

[13] Considera-se que a disciplina *Inteligência Artificial* passa a existir enquanto campo de investigação reconhecido pela comunidade científica a partir da chamada *Dartmouth Summer Research Project on Artificial Intelligence*, em 1956. Essa conferência contou com a participação daqueles que seriam mais tarde os expoentes do campo: John McCarthy, idealizador da conferência e criador do nome da disciplina, Marvin Minsky, Calude Shannon, Oliver Selfridge, Nathaniel Rochester, entre outros. A disciplina *Lingüística Computacional*, por sua vez, cujo nome foi cunhado em 1967 por David Hays (Moreno Fernández, 1990, p. 6), antes de se especializar enquanto uma disciplina que focaliza alguns aspectos do estudo computacional das línguas naturais, focalizava essencialmente o estudo das linguagens formais e das linguagens de programação.

Petrick (1990) sinaliza uma influência marcante dos estudos computacionais desenvolvidos no âmbito do PLN sobre o desenvolvimento da Teoria Lingüística. Destaque especial merecem as investigações que vieram reanimar a discussão sobre as propriedades formais das gramáticas das línguas naturais (Harlow & Vincent, 1989). A utilização de *gramáticas sintagmáticas livres de contexto ampliadas* como modelo de descrição lingüística e a crítica de Gazdar (1982) a alguns aspectos da obra chomskiana atestam essa retomada.

Os argumentos que Chomsky (1957) construiu para demonstrar que as gramáticas sintagmáticas livres de contexto eram inadequadas para caracterizar a sintaxe das línguas naturais perderam sua força com a proposição das redes de transição ampliadas de Woods (1970). Tomando por base o fato de os falantes processarem as estruturas lingüísticas instantaneamente e os resultados obtidos com a implementação computacional da "Teoria Padrão" (Chomsky, 1965), Gazdar (1982) mostra que as gramáticas gerativas, com seu grande número de dispositivos formais, são completamente inadequadas para servir de modelo de processamento das estruturas lingüísticas pelos falantes. A partir dessa análise, constrói um novo modelo de gramática, sem as clássicas "transformações" e o denomina Gramática Sintagmática Generalizada, origem da gramática *Head-driven Phrase Structure Grammar*, mais conhecida pela sigla *HPSG* (Pollard & Sag, 1994). Esses dois exemplos evidenciam que a argumentação de base computacional pode trazer novos recursos para se repensar as teorias lingüísticas.

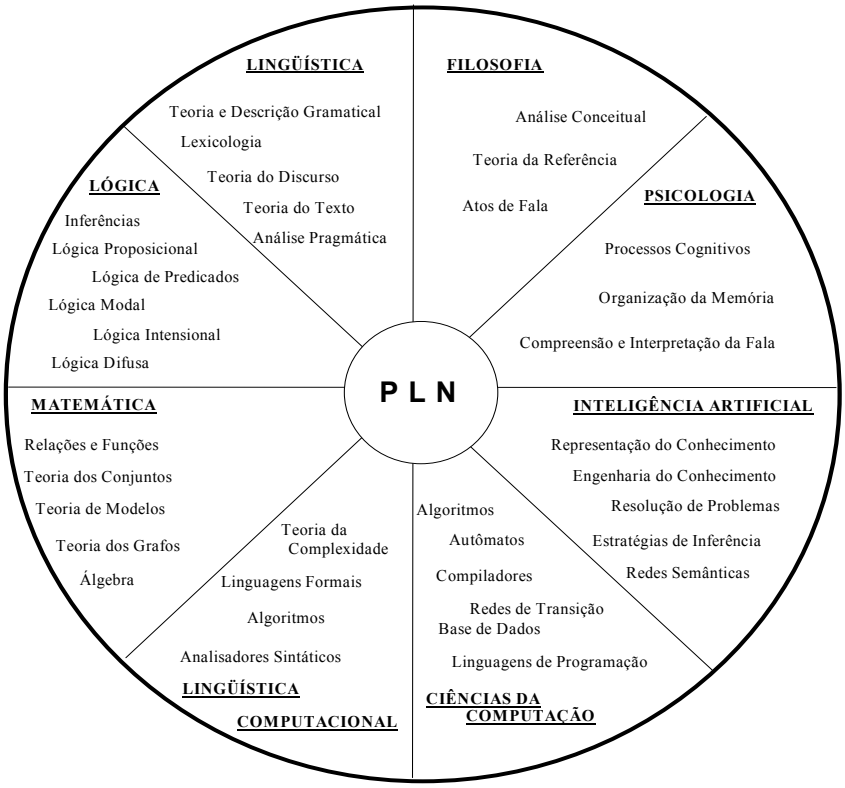
Além disso, o estudo do PLN sobrepõe-se a parcelas dos domínios da Filosofia da Linguagem, Lingüística e Psicologia, ao procurar compreender, com sua lente, a linguagem humana, suas funções, sua manifestação nas diferentes línguas, sua estrutura interna e sua relação com a realidade, com os processos de raciocínio e com o comportamento verbal.

Sobrepõe-se também a domínios da Lógica, Matemática, Ciências da Computação, Lingüística Computacional e Inteligência Artificial, ao procurar, nessas disciplinas, instrumentos e estratégias indispensáveis à construção e à implementação dos modelos. São exemplos: sistemas de representações formais, como o cálculo de predicados, a lógica modal e temporal, os grafos de representação lexical, sintática, semântica e conceitual, as linguagens de programação, os autômatos, as gramáticas e os algoritmos de análise gramatical; os sistemas de representação do conhecimento de mundo e de crenças; as estratégias de resolução de problemas e de organização da informação.

Há que se observar, ainda, que existem sobreposições entre a Lingüística, de um lado, e a Filosofia, a Lógica, a Matemática, as Ciências da Computação e a Inteligência Artificial, de outro. A mais clássica delas é a sobreposição que se constata entre os Lingüística, Filosofia e Lógica (Frege, 1990; Austin, 1962; 1990; Grice, 1990; Searle, 1990a e 1990b; Reichenbach, 1947; Lehmann et al., 1985; Barwise & Perry, 1983; 1990). Lembre-se de que aquela originou-se no seio destas disciplinas: a tradição gramatical do ocidente encontrou, nessas áreas clássicas, um dos modelos mais profícuos para o estudo das categorias e funções gramaticais, do conteúdo semântico das orações e das atitudes proposicionais (Allwood, 1977; McCawley, 1981).

Outras importantes contribuições mútuas também merecem destaque (Meulen, 1989). De um lado, constata-se o "viés lingüístico" que tomou conta da filosofia analítica de Oxford e Cambridge. A análise conceitual desenvolvida por essa corrente filosófica, sob a denominação de "análise componencial", acabou por ser aplicada aos estudos de decomposição do significado dos itens lexicais. Com a denominação de "o estudo do significado em uso", sob a influência de Austin e Wittgenstein, transformou-se em um método de análise que procura investigar como uma expressão lingüística pode ter significados diferentes em diferentes contextos de uso, enfatizando, portanto, a forte dependência contextual do significado. De outro, com o desenvolvimento da lógica moderna, por lógicos como Frege e Russell, e com a aplicação dos seus métodos e resultados aos estudos lingüísticos, temas como "extensão e intensão", "contextos transparentes e opacos" e "interpretação *de re* e *de dicto*" passaram a fazer parte das discussões sobre a semântica das línguas naturais.

Figura 3. Recursos teórico-metodológicos para o estudo do PLN.



Já a Matemática, as Ciências da Computação e a Inteligência Artificial só muito recentemente passaram a instrumentalizar a Lingüística no equacionamento de alguns dos seus problemas cruciais: os modelos que servem de base para a descrição sintática das teorias lingüísticas modernas se utilizam de regras de produção, de grafos e de funções matemáticas; estruturas como *frames*, *scripts* e *plans* (Schank & Abelson, 1977; Dyer et al., 1990), provenientes dos estudos de Inteligência Artificial são empregados por teorias semânticas e do discurso.

Na Figura 3, classificados segundo as disciplinas matrizes, apresento a sistematização dos principais recursos teórico-metodológicos de que o estudo do PLN dispõe.

O estudo do PLN é um domínio de pesquisa privilegiado, amplo e fecundo, uma vez que a construção do corpo de conheci-

mentos necessários para a implementação de sistemas de PLN, com o grau de sofisticação delineado neste trabalho, exige seleção, organização, representação e codificação de uma variedade de informações na complexa tarefa de criar um simulacro computacional da competência e do desempenho lingüísticos humanos.

Referências

AKMAJIAN, A. et al. *Linguistics: an introduction to language and communication*. Cambridge, Mass.: The MIT Press, 1986.

ALLEN, J. F. *Natural language understanding*. Menlo Park: Benjamin Cummings, 1987.

ALLEN, J. F. & PERRAULT, C. R. Analyzing Intentions in Utterances. *Artificial intelligence*, 15, p. 143-78, 1980.

ALLWOOD, J. et al. *Logic in linguistics*. Cambridge: Cambridge University Press, 1977.

AMAREL, S. Problem Solving. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 767-79, 1990.

APPELT, D. E. *Planning English sentences*. Cambridge: Cambridge University Press, 1985.

AUSTIN, J. L. *How to do things with words*. Oxford: Oxford University Press, 1962.

———. Performative Utterances. In: A. P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 105-14, 1990.

BALLARD, B. W.; JONES, M. A. Computational Linguistics. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 133-51, 1990.

BARTON, G. E., BERWICK, R. C.; RISTAD, E. S. *Computational complexity and natural language*. Cambridge, Mass.: The MIT Press, 1987.

BARWISE, J.; PERRY, J. *Situations and attitudes*. Cambridge, Mass.: The MIT Press, p. 27-45, 1983.

———. Semantic Innocence and Uncompromising Situations. In: A. P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 392-404, 1990.

BIERMANN, A. Automatic Programming. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 18-35, 1990.

BORBA, F. S. *Introdução aos estudos lingüísticos*. 8. ed. São Paulo: Ed. Nacional, 1984.

BRACHMAN, R. J.; LEVESQUE, H. J. *Readings in knowledge representation*. San Mateo: Morgan Kaufmann, 1985.

BRESNAN, J. An Approach to Universal Grammar and The Mental Representation of Language. *Cognition*, **10**, p. 39-52, 1981.

——— (ed.). *The mental representation of grammatical relations*. Cambridge, Mass.: The MIT Press, 1982.

CARBONELL, J. G.; HAYES, P.J. Natural-Language Understanding. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 660-77, 1990.

CARPENTER, R.; THOMASON, R. Inheritance Theory and Path-Based Reasoning: An Introducton. In: H. E. Kyburg et al. (eds.). *Knowledge representation and defeasible reasoning*. Dordrecht: Kluwer, p. 309-43, 1990.

CHOMSKY, N. *Syntactic structures*. Haia: Mouton, 1957.

——— . *Aspects of the theory of syntax*. Cambridge, Mass.: The MIT Press, 1965.

——— . *Knowledge of language: its nature, origins, and use*. New York: Praeger, 1986.

COHEN, P. R. & PERRAULT, C. R. Elements for A Plan-Based Theory of Speech Acts. *Cognitive Science*, **3**, p. 177-212, 1979.

CLOCKSIN, W. F.; MELLISH, C. S. *Programming in prolog*. Berlin: Springer-Verlag, 1987.

DIAS-DA-SILVA, B. C. *O fenômeno da apassivação: em busca da passiva protótipo*. Araraquara, 160 p. Dissertação (Mestrado em Letras) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, 1990.

DIK, S. *The theory of functional grammar*. Part 1 and Part 2. Berlin: Mouton de Gruyter, 1997.

DOWTY, D. R. et al. *Introduction to Montague semantics*. Dordrecht: Reidel, 1981.

——— . et al. *Natural language parsing*. Cambridge, Mass.: The MIT Press, 1985.

DYER, M.G. et al. Scripts. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 980-94, 1990.

EARLEY, J. An Efficient Context-Free Parsing Algorithm. *Communications of The Association for Computing Machinery*, 14, p. 453-60, 1970.

FÁVERO, L. L. E.; KOCH, I. G. V. A Lingüística Textual. In: ——— *Lingüística textual: introdução*. São Paulo: Cortez, p. 11-25, 1983.

FODOR, J. D.; FRAZIER, L. Is The Human Sentence Parsing Mechanism an ATN. *Cognition*, **8**, p. 417-59, 1980.

FRAZIER, L. & FODOR, J. The Sausage Machine: A New Two Stage Parsing Model. *Cognition*, **6**, p. 291-325, 1978.

FREGE, G. On Sense and Nominatum. In: A. P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 190-202, 1990.

GARRET, M. F. Sentence Processing. In: D. N. Osherson & H. Lasnik (eds.). *An invitation to cognitive science: language*. Cambridge: Mass., The MIT Press, p. 133-75, 1990.

GAZDAR, G. Phrase Structure Grammar. In: P. Jacobson & G. K. Pullum (eds.). *The nature of syntactic representation*. Dordrecht: D. Reidel, p. 131-86, 1982.

GAZDAR, G.; MELLISH, C. (1989). *Natural language processing in prolog: an introduction to computational linguistics*. New York: Addison-Wesley, 1989.

GREIMAS, A. J.; COURTÉS, J. *Dicionário de semiótica*. São Paulo: Cultrix, 1979.

GRICE, H. P. Logic and Conversation. In: A. P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 149-60, 1990.

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

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

HALLIDAY, M. A. K.; HASAN, R. *Cohesion in English*. London: Longmans, 1976.

HALVORSEN, P.-K. Computer Applications of Linguistic Theory. In: F. Newmeyer (ed.). *Linguistics: the cambridge survey II*. Cambridge, Mass.: Cambridge University Press, p. 198-219, 1989.

HARLOW, S.; VINCENT, N. Generative linguistics: an overview. In: F. Newmeyer (ed.). *Linguistics: the cambridge survey II*. Cambridge: Cambridge University Press, p. 1-17, 1989.

HAYES-ROTH, F. Expert Systems. In: E. Shapiro (ed.) . *Encyclopedia of artificial intelligence*. New York: Wiley, p. 287-98, 1990.

HEARN, A. C. et al. Computational Linguistics. In: B. W. ARDEN (ed.) *What can be automated*. Cambridge, Mass.: The MIT Press. p. 538-48, 1980.

HIRST, G. *Semantic interpretation and the resolution of ambiguity*. Cambridge, Mass.: Cambridge University Press, 1992.

HOBBS, J. R. et al. *Interpretation as Abduction*. Technical Note 499, Artificial Intelligence Center, SRI International, Menlo Park, California, 1990.

HORN, L. R. Pragmatic theory. In: F. Newmeyer (ed.). *Linguistics: the cambridge survey I*. Cambridge: Cambridge University Press, p. 113-45, 1988.

JACKENDOFF, R. *Semantic structures*. Cambridge: Mass., The MIT Press, 1990.

JAKOBSON, R. *Lingüística e comunicação*. São Paulo: Cultrix, 1977.

KAY, M. Parsing in Functional Unification Grammar. In: D. R. Dowty et al. (eds.). *Natural language parsing*. Cambridge: Cambridge University Press, p. 251-78, 1985.

KEMPSON, R. (1977). Grammar and conversational principles. In: F. Newmeyer (ed.). *Linguistics: the cambridge survey I*. Cambridge: Cambridge University Press, p. 139-63, 1988.

KLAVANS, J. Computational Linguistics. In: O'GRADY, W. et al. *Contemporary linguistics*. New York: St. Martin's Press, cap. 15, p. 413-47, 1989.

KORFHAGE, R. R. *Logic and algorithms*. New York: Wiley, 1966.

KRONFELD, A. *Reference and computation*. Cambridge: Cambridge University Press, 1990.

LASNIK, H. The Study of Cognition. In: D. N. Osherson & H. Lasnik, (eds.). *Language: an Invitation to cognitive science*. Cambridge, Mass.: The MIT Press, p. xi-xix, 1990.

LEECH, G. *Principles of pragmatics*. London: Longman, 1983.

LEHMANN, W. P. et al. Human Language and Computers. *Computers and The Humanities*, **19**, p. 77-83, 1985.

LEMLE, M. *Análise sintática: teoria geral e descrição do português*. São Paulo: Ática, 1984.

LEVINSON, S. C. *Pragmatics*. Cambridge: Cambridge University Press, 1983.

LYONS, J. *As idéias de Chomsky*. São Paulo: Cultrix, 1976.

———. *Semantics 1& 2*. London: Cambridge University Press, 1977.

———. *Introdução à lingüística teórica*. São Paulo: Cia. Ed. Nacional-EDUSP, 1979.

———. *Linguagem e lingüística*. Rio de Janeiro: Zahar, 1981.

MARANTZ, A. *On the nature of grammatical relations*. Cambridge, Mass.: The MIT Press, 1984.

MARCUS, M. P. *A theory of syntactic recognition for natural language*. Cambridge, Mass.: The MIT Press, 1980.

McCAWLEY, J. D. *Everything that linguists have always wanted to know about logic*. Chicago: The University of Chicago Press, 1981.

McKEOWN, K. R. *Text generation*. Cambridge: Cambridge University Press, 1985.

MEULEN, A. ter Linguistics and The Philosophy of Language. In: F. Newmeyer (ed.). *Linguistics: the Cambridge survey I*. Cambridge: Cambridge University Press, p. 430-46, 1989.

MILLER, G. A., FELLBAUM, C. Semantic networks of English. *Cognition*, 41, n. 1-3, p. 197-229, 1991,

MINSKY, M. *Semantic information processing*. Cambridge, Mass.: MIT Press, 1968.

———. A Framework for Representing Knowledge. In: J. Haugeland (ed.). *Mind design*. Cambridge, Mass.: The MIT Press, p. 95-128, 1975.

MORENO FERNÁNDEZ, F. Lingüística Informática e Informática Lingüística. *Lingüística Española Actual*, **12**, p. 5-16, 1990.

MYKOWIECKA, A. Natural-language generation – an overview. *International Journal of Man-Machine Studies*, **34**, p. 497-511, 1991.

NIRENBURG, S. et al. *Machine translation*. San Mateo: Morgan Kaufmann, 1992.

PARTEE, B. H. et al. *Mathematical methods in linguistics*. Dordrecht: Kluwer, 1993.

PEREIRA, F. C. N.; SHIEBER, S. *Prolog and natural language analysis*. Chicago: The University of Chicago Press, 1987.

PERRAULT, C. R. On The Mathematical Properties of Linguistic Theories. *Computational Linguistics*, **10**, 3-4, p. 165-76, 1984.

PETRICK, S. Parsing. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, p. 687-96, 1990.

POLLARD, C.; SAG, I. *Head-driven phrase structure grammar*. Chicago: CSLI Publications, 1994.

PRINCE, E. F. Discourse analysis: a part of the study of linguistic competence. In: F. Newmeyer (ed.). *Linguistics: the Cambridge survey I*. Cambridge: Cambridge University Press. p. 164-82, 1988.

PRITCHETT, B. L. Garden Path Phenomena and The Grammatical Basis of Language Processing. *Language*, **64**, p. 539-76, 1988.

PUSTEJOVSKY, J. *The generative lexicon*. Cambridge, Mass.: The MIT Press, 1995.

PUSTEJOVSKY, J.; BOGURAEV, B. Lexical Knowledge Representation and Natural Language Processing. *IBM Journal of Research and Development*, **35**, p. 1-20, 1991.

QUILLIAN, M. R. Semantic Memory. In: M. Minsky. *Semantic information processing*. Cambridge, Mass.: MIT Press. p. 227-70, 1968.

REICHENBACH, H. *Elements of symbolic logic*. New York: Macmillan, 1947.

REYTER, R. Nonmonotonic Reasoning." *Annual Review of Computer Science*, **2**, p. 147-86, 1987.

RICH, E. (1983). *Inteligência artificial*. Trad. N.Vasconcellos. (Rev. técnica Nizam Omar). São Paulo: McGraw-Hill, 1983.

———. Artificial Intelligence and The Humanities. *Computers and The Humanities*, **19**, p. 117-22, 1985.

SADOK, J. M. Speech act distinctions in grammar. In: F. Newmeyer (ed.). *Linguistics: the cambridge survey I*. Cambridge: Cambridge University Press, p. 183-97, 1988.

SANDERS, A.; SANDERS, R. Syntactic Parsing: A Survey. *Computers and The Humanities*, **23**, p. 13-30, 1989.

SCHANK, R. C.; ABELSON, R. *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum, 1977.

SCHANK, R. C.; RIESBECK, C. K. (eds.) *Inside computer understanding*. Hillsdale, NJ: Lawrence Erlbaum, 1981.

SEARLE, J. R. What Is a Speech Act? In: A.P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 115-25, 1990a.

———. Indirect Speech Acts. In: A. P. Martinich (ed.). *The philosophy of language*. Oxford: Oxford University Press, p. 161-75, 1990b.

SELLS, P. *Lectures on contemporary syntactic theories*. Chicago: The University of Chicago Press, 1985.

SHAPIRO, S. C. Bottom-up and top-down processing. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley. p. 779-85, 1990.

SLAGLE, J.; GINI, M. Pattern Matching. In: E. Shapiro (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley. p. 716-20, 1990.

STAROSTA, S. Natural Language Parsing and Linguistic Theories: Can The Marriage Be Saved? *Studies in Language*, **15**, p. 175-97, 1991.

STERLING, L.; SHAPIRO, E. *The art of prolog*. Cambridge, Mass.: The MIT Press, 1986.

SUDKAMP, T. A. *Languages and machines*. Reading, Mass.: Addison-Wesley, 1991.

VARILE, G. B.; ZAMPOLLI, A. *Survey of the state of the art in human language tecnology*. Cambridge: Cambridge University Press, 1997.

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

WINSTON, P. H. *Artificial intelligence*. Reading, Mass.: Addison-Wesley, 1984.

WOODS, W. A. Transition Network Grammars for Natural Language Analysis. *Communications of The Association for Computing Machinery*, **13**, p. 591-6, 1970.