UNIVERSIDADE ESTADUAL PAULISTA

FACULDADE DE CIÊNCIAS E LETRAS

CÂMPUS DE ARARAQUARA

A FACE TECNOLÓGICA DOS ESTUDOS DA LINGUAGEM: o processamento automático das línguas naturais

Tese apresentada para obtenção do Título de DOUTOR em LETRAS – na área de concentração Lingüística e Língua Portuguesa – à Faculdade de Ciências e Letras da Universidade Estadual Paulista, sob a orientação do *Prof. Dr. Telmo Correia Arrais*.

Por: ***Bento Carlos Dias da Silva***

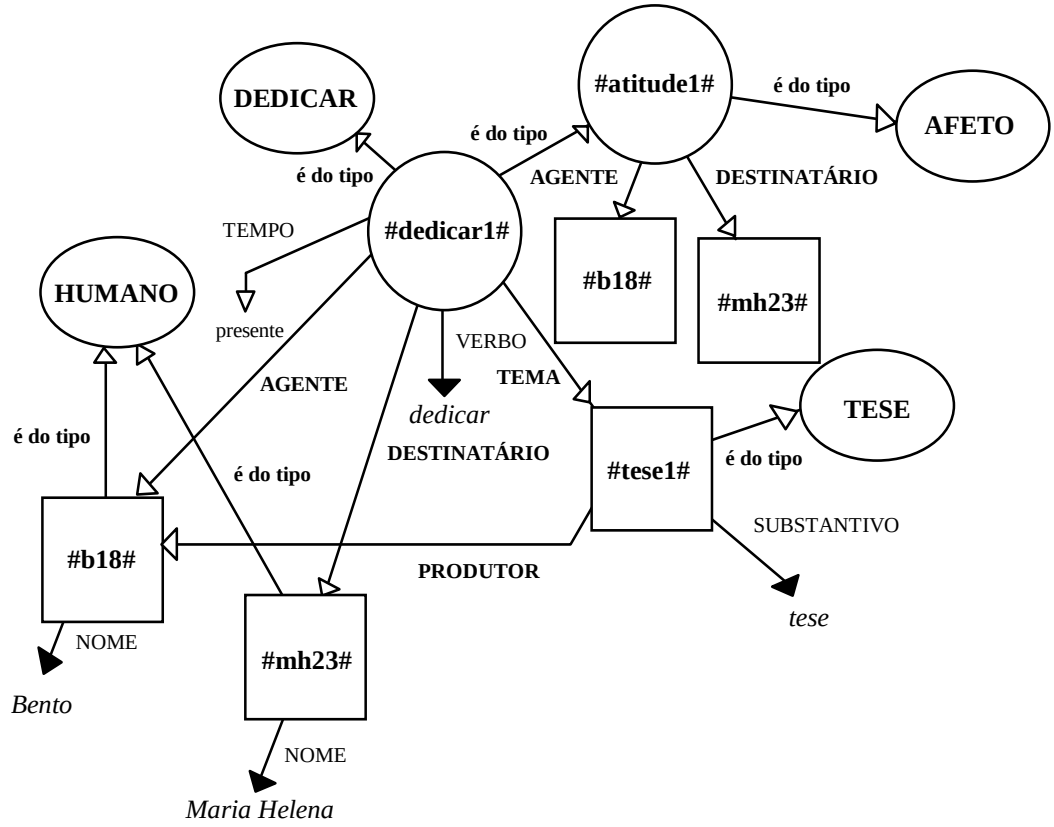
A R A R A Q U A R A

1 9 9 6

BENTO CARLOS DIAS DA SILVA

A FACE TECNOLÓGICA DOS ESTUDOS DA LINGUAGEM:
o processamento automático das línguas naturais

A R A R A Q U A R A
1 9 9 6



(estrutura texto
(orações (valor oração_1))
(relações (valor relação_1))
(atitudes (valor atitude_1))
(intenções_do_locutor (valor intenção_do_locutor1)))
(estrutura oração_1
(núcleo (valor **#dedicar1#**))
(aspecto (duração prolongada)
(fase início)
(iteração_1))
(tempo (valor tempo_1)))
(estrutura **#dedicar1#**
(é (valor (**DEDICAR**))
(agente (valor **#b18#**))
(destinatário (valor **#mh23#**))
(tema (valor **#tempo1#**)
(estrutura …

Bento dedica a tese à Maria Helena

Agradecimentos

Decifrar a esfinge é um trabalho solitário. Algumas pessoas e instituições, entretanto, forneceram base segura para que eu pudesse enfrentar esse desafio. A elas, meu reconhecimento público.

Ao **Prof. Dr. Telmo Correia Arrais,** orientador da tese, pela confiança e incentivo constantes.

Aos professores do Programa de Pós Graduação em Lingüística Computacional da Carnegie Mellon University, em Pittsburgh/EUA, sobretudo, **David A. Evans**, **Lori Levin** e **Brad Pritchett**, por me iniciarem nesta área tão controvertida.

Aos colegas de Departamento de Letras Modernas da Faculdade de Ciências e Letras da UNESP de Araraquara e às Áreas de Língua e Literatura Inglesa e Norte Americana, sobretudo **Leila Cury Rodrigues Olivi** e **Ademar da Silva**, por terem assumido um ônus maior de trabalho didático, para que meu envolvimento com esta tese fosse possível.

À **Profª. Drª. Sonia Veasey Rodrigues**, pela confiança e estímulo, sempre otimista.

Ao **Prof. Dr. Francisco da Silva Borba** e à **Profª. Drª. Beatriz Nunes de Oliveira Longo**, pela inestimável contribuição que trouxeram ao trabalho quando da realização do exame de qualificação.

À **Profª. Drª Lídia Fachin**, pelo *résumé*.

À **Profª. Thereza Anália Cochar Magalhães**, pela revisão dos originais.

Agradecimento especial dedico aos funcionários da **FCL**, sobretudo, aos meus companheiros do **Departamento de Letras Modernas**, da **Biblioteca**, da **Pós Graduação** e do **Pólo Computacional**, parceiros sempre solidários e solícitos nessa trajetória.

À **minha família** e aos **meus amigos**, agradeço a paciência.

Agradecimento final à **CAPES** (Coordenadoria de Aperfeiçoamento de Pessoal do Ensino Superior), com a certeza de que, sem o financiamento para o estágio de um ano nos Estados Unidos, a esfinge teria continuado inacessível para mim.

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.

R E S U M O

O campo de estudos do processamento automático de línguas naturais (PLN) apresenta um crescimento surpreendente e uma heterogeneidade de projetos que se espalham desordenadamente. Além disso, reúne pesquisadores com embasamentos téoricos e interesses bastante diversos, enfatizando diferentes aspectos dos problemas e empregando uma pluralidade de métodos e técnicas. Nele, não é difícil apontar falhas. Entre elas, encontra-se o não raro tratamento superficial dado aos complexos fenômenos da linguagem, fato que evidencia a incômoda lacuna que separa os estudos sobre o PLN e a teoria lingüística. Incentivar a integração dessas duas áreas, porém, não é tarefa trivial. Além da própria complexidade dos fenômenos lingüísticos, é preciso também enfrentar os desencontros e a massiva quantidade de relatórios de pesquisa, artigos e resenhas, contendo uma multiplicidade de enfoques e de formalismos intrigantes.

Diante da ausência de trabalhos solidários e, principalmente, reconhecendo a importância de contribuições mútuas que podem passar a existir entre os dois domínios, esta tese tem por objetivo enfrentar esses e outros problemas na tentativa de minimizar o divórcio entre os "cientistas da linguagem" e os "engenheiros da linguagem". Em particular, enfatiza a necessidade de trabalho cooperativo, envolvendo lingüistas e projetistas de sistemas de PLN, dando destaque ao potencial interdisciplinar, científico e tecnológico resultante dessa parceria. Assim, nela, caracteriza-se uma abordagem que incentiva o trabalho cooperativo entre os diversos especialistas envolvidos e delineia-se um quadro de referência para alunos e pesquisadores das Humanidades, cujas preocupações incluem a investigação das línguas naturais do ponto de vista computacional.

Nesse sentido, propõe-se que um sistema de PLN seja considerado um tipo particular de "sistema de processamento automático de conhecimentos", em que um complexo de informações lingüísticas e extralingüísticas é representado e automaticamente aplicado na investigação ou execução de tarefas que envolvem conhecimentos de natureza lingüística: revisão ortográfica, construção de gramáticas e léxicos, tradução e sumarização automáticas, interpretação e produção de textos. Assume-se, portanto, que o programa de pesquisa sobre o

PLN deve espelhar os métodos e técnicas propostos para o desenvolvimento desse tipo de sistemas de conhecimento.

Em linhas gerais, argumenta-se que tanto o programa de pesquisa como os projetos de desenvolvimento de sistemas de PLN específicos precisam enfrentar os níveis de processamento gramatical e pragmático-discursivo em três domínios: o Lingüístico, o Representacional (Linguístico-computacional) e o Implementacional (Computacional). Do ponto de vista operacional, propõem-se, então, estas três fases de análise solidárias, cíclicas e progressivas: (i) Fase de representação lingüística (conceituação do objeto lingüístico a ser modelado), em que se analisam as parcelas do conhecimento e desempenho lingüísticos que serão incorporadas ao sistema; (ii) Fase de representações e algoritmos (modelagem da conceituação), que transforma os objetos lingüísticos descritos na fase anterior em representações formais, contendo todas os elementos e as especificações conceituais do sistema; e (iii) Fase da contrução do sistema (implementação do modelo), que codifica as representações propostas na fase anterior em programas de computador e projeta os diferentes componentes do sistema, tais como as interfaces usuário-máquina, operacional e para o desenvolvimento do próprio sistema.

DIAS-DA-SILVA, B.C. (1996) ***The technological facet of language studies: natural language processing***. Doctoral Dissertation, Faculdade de Ciências e Letras, UNESP (São Paulo State University), Araraquara / SP, BRAZIL. 272p.

A B S T R A C T

It is a fact that an overwhelming growth in the field of natural language processing has taken place. But it is also a fact that a multiplicity of natural language processing projects has sprawled. As a result, natural language processing seems to be a discipline in ferment, which gathers researchers with a wide range of backgrounds and interests, emphasizing its diverse aspects and employing manifold methods and techniques. Thus, despite the enthusiasm, there have been some drawbacks, some of which due to either lack of appreciation for the complexity of natural languages or underspecification of the complexity of the task itself. Furthermore, there has been a disturbing gap between natural language processing and linguistic theory. On the one hand, it is not difficult to spot natural language processing projects that either resort to inappropriate linguistic models or strive to succeed without any recourse to linguistic theory; on the other hand, linguistic theory has either disregarded computational issues completely or provided the ammunition to deaden the enthusiastic development of natural language computer applications. In addition, those who are new to either field have to confront an astounding number of technical reports, journal articles and conference papers, to get acquainted with a number of approaches, and to decode puzzling formalisms.

Given such a lack of team work, this dissertation aims to confront these and other problems in an attempt to reduce the gap between "language scientists" and "language engineers". In particular, it stresses the need for cooperative work between linguists and natural language processing system designers, and emphasizes the task of developing natural language processing projects that are well-defined and linguistically motivated as well as its interdisciplinary, scientific, and technological potential. Accordingly, it characterizes an approach to natural language processing that fosters cooperative work between different specialists and attempts to present a unified framework to students and researchers in linguistics and related disciplines whose concerns include tackling the fascinating computer approach to the understanding of natural languages.

In order to accomplish these goals it is suggested that a natural language processing system is a particular type of knowledge processing system where a complex of linguistic and extra-linguistic knowledge is represented and applied electronically to exploit and to perform natural language tasks such as spelling

checking, grammar and lexicon building, machine translation, and natural language understanding and generation. Thus it is assumed that natural language processing research programs should mirror the knowledge processing system development strategies.

Accordingly, it is claimed that both the natural language processing research program and the task of building particular natural language processing systems should tackle the grammatical and discourse phases of processing in three broad domains – Linguistic, Representational (Computational-linguistic), and Implementational (Computational) domains. Three iterative and evolutionary phases of analysis are posited for both: (i) Phase of linguistic representations (conceptualization), which provides descriptions of both knowledge of language and language use, (ii) Phase of computer representations and algorithms (formalization), which abstracts from the previous phase to describe how linguistic objects are to be computationally encoded, and provides additional knowledge representations, and (iii) Phase of system building (implementation), which abstracts even further to provide computer programs and system components such as user interfaces, development and operational environments.

DIAS DA SILVA, B.C. ***La face technologique des études du langage: l'analyse automatique des langues naturelles***. Faculdade de Ciências e Letras/UNESP, Araraquara/SP, BRÉSIL, 1996. (Thèse). 272p.

R É S U M É

Le champ des études concernant l'analyse automatique des langues naturelles (PLN) présente une croissance surprenante et une héterogéneité de projets se répandant de façon désordonnée.En plus, il rassemble des chercheurs avec des fondements théoriques et des intérêts fort diversifiés, mettant l'accent sur différents aspects des problèmes et se servant d'une pluralité de méthodes et de techniques. Il n'est pas difficile d'en montrer les graves défauts, dout, très souvent, le traitement superficiel consacré aux phénomènes complexes du langage, ce qui rend évidente la lacune gênante qui écarte les études sur le PLN de celles de la théorie linguistique.

Stimuler l'íntégration des ces deux domaines ne constitue pourtant pas une tâche facile. En plus, de la complexité même des phénomènes linguistiques, il faut affronter également une série de mesententes et surtout la quantité massive de rapports de recherches, d'articles et des comptes rendus appuyés sur une multiplicité d'approches et de formalismes bizarres.

Devant l'absence de travaux solidaires et surtout reconnaissant l'importance de contribuitions mutuelles qui pourront s'etablir entre ces deux domaines, cette thèse a pour but de faire face à toutes ces questions et à d'autres pouvant apparaître das la suite, et essayer de réduire à de moindres proportions le divorce entre les "scientifiques du langage" et les "ingénieurs du langage". Surtout cette thèse met l'accent sur la necessité du travail coopératif comprenant des linguistes et des chercheurs qui sont en mesure de projeter des systèmes pour PLN, surtout en ce qui concerne les potentialités interdisciplinaires, scientifiques et technologiques résultant de cette association.Ainsi se caractérise une approche stimulant le travail coopératif entre les différents spécialistes travaillant ensemble et un repère s'ebanche pour étudiants et chercheurs des Sciences Humaines dont une des préoccupations est d'explorer les langues naturelles du point de vue de l'ordinateur.

En ce sens, il est suggéré qu'un système de PLN est une sorte particulière de "système d'analyse automatique des connaissances"ou un complexe d'informations linguistiques et extralinguistiques est représenté et automatiquement appliqué à l'investigation ou à l'exécution de tâches concernant des connaissances de nature linguistique: révision orthographique, élaboration de grammaires et de lexiques, traduction automatique, interprétation et génération automatique de textes. Il

devient done évident que le programme de recherches sur le PLN doit réflécher les méthodes et les techniques proposés préalablement en vue du développement de ce genre de système des coinnaissances.

On argumente, dans cette thèse, que le programme de recherches aussi bien que les projets de développement des systèmes spécifiques de PLN doivent être confrontés aux niveaux d'analyse grammaticale et pragmatico-discursif dans trois domaines: linguistique, de représentation et d'implémentation. Du point de vue opérationnel, son propose donc trois phases solidaires, cycliques et progressives d'analyse: (i) Phase de représentation linguistique (définition) , qui analyse les parcelles de connaissance et de performance linguistiques qui seront incorporées dans le système, (ii) Phase de représentations et d'algorithmes (formalisation) , qui transforme les objets linguistiques décrits dans la phase précédente en représentations formelles, qui, elles, contiennent les éléments et les spécifications conceptuelles du système et (iii) Phase du construction du systéme (implémentation), qui codifie dans des programmes les représentations proposées par la phase précédente et projette les différentes composantes du système, telles qu'une interface usager-machine, une interface opérationnelle et une interface pour le développement du système lui-même.

Sumário

Prefácio

> *"Wouldn't it be nice just to sit down at some computer terminal and tell the computer, in whatever language you speak, some task that you want done and have the computer do it?"*
>
> Rachel Reichman (1985: xi)

Meu interesse por pesquisas lingüísticas, em especial por "pesquisas lingüísticas computacionalmente motivadas", é resultado de duas atividades, para muitos irreconciliáveis, que passei a desenvolver desde a minha primeira graduação em Matemática, na década de 70: explorar possibilidades de aplicação de recursos computacionais a outras áreas do conhecimento e estudar a língua inglesa.

Entre as primeiras investigações, dois fenômenos lingüísticos passaram a ser o centro das minhas preocupações: a correlação sintática e a identidade semântica entre pares de orações ativas e passivas, observadas nas línguas em geral, e as orações "passivas peculiares" do inglês, conhecidas também como "orações passivas oblíquas".

Estudava, na ocasião, a possibilidade de construção de um programa de computador (já que havia dado os primeiros passos como programador), capaz de transformar orações ativas em orações passivas, e vice-versa. Tentava, sem o saber, buscar uma solução computacional para o problema colocado pelo processamento automático das orações passivas. No entanto, à medida que tentava aprimorar o programa para o processamento dos mais variados tipos de construções passivas do

inglês, mais problemas iam surgindo. Quantos tipos de construções passivas existem? Como isolar os elementos relevantes para escrever o programa? Como escrever um programa semelhante para o português?

A busca de respostas para questões como essas me levou, primeiro, para o estudo das Linguagens Formais, então emergente, e, posteriormente, para a Lingüística, disciplina que era completamente desconhecida para mim. Ao ler o livro *Logic and Algorithms*, de Robert R. Korfhage (1966), que dedica um pequeno capítulo à teoria das linguagens formais, constatei o papel decisivo que Noam Chomsky desempenhou para o desenvolvimento dessa teoria.

Dei então o primeiro passo. Estudei o livro *O que é lingüística? Uma introdução ao pensamento de Noam Chomsky*, de John Lyons (1976), uma vez que a leitura do trabalho original *Syntactic Structures* de Chomsky (1957) exigia essa contextualização. Lyons revelou-me, para a minha satisfação, que as matemáticas e o estudo das línguas poderiam usufruir um do outro. Eureca! Existiam pesquisas na "Área de Humanas" que, de fato, eram "computacionalmnte motivadas"!

Informalmente, começava a tecer os primeiros fios de um complexo elo entre dois domínios do conhecimento, aparentemente desconexos: as Humanidades e as Matemáticas. Ou, como se costuma dizer: entre as "Exatas" e as "Humanas".

Infelizmente, conhecer um pouco das ciências matemáticas não me autorizou a desenvolver pesquisas lingüísticas. Precisava estar oficialmente inserido no "Mundo das Humanidades"… Afinal, o que um matemático poderia entender de lingüística?

Em 1981, então, mais uma graduação me aguardava… Desta vez: Letras. Graduação psicologicamente custosa, porque,

enquanto colegas e amigos rumavam ao Mestrado e definiam suas carreiras profissionais, lá estava eu – "o velho" – retornando à graduação como intruso. Enfim, mais quatro anos…

Terminada a graduação em Letras, a tão esperada pós-graduação…

Na Dissertação de Mestrado, *O fenômeno da apassivação: em busca da passiva protótipo* (DIAS-DA-SILVA, 1990), registrei as minhas primeiras reflexões sobre esse fenômeno da linguagem. Nela, pude apreciar um pouco da história dos estudos sobre a apassivação, "lutei" com modelos de análise lingüística alternatives e até conflitantes entre si e procurei sistematizar um conjunto significativo de estruturas e de atualizadores da construção passiva. Como conclusão, apresentei uma possível caracterização desse fenômeno em termos de uma "Passiva Protótipo": uma representação da forma gramatical e da função pragmático-discursiva prototípicas das construções passivas.

A seguir, dei prosseguimento ao meu projeto acadêmico. Somei esforços e, com o incentivo de professores, colegas e amigos, visitei universidades americanas em busca de um programa de pós-graduação que não só investisse em pesquisas sobre o processamento automático de línguas naturais como também valorizasse a integração de pesquisas lingüísticas e computacionais, uma vez que o Brasil não dispunha de programas com esse perfil. Os contatos com pesquisadores, que tive a oportunidade de estabelecer durante a visita, permitiram que eu conhecesse um dos raros programas de pós-graduação que se aproximava do perfil procurado: o Programa de Pós-Graduação em Lingüística Computacional da Universidade Carnegie Mellon, EUA.

No início de 1991, motivado por essas experiências promissoras e com o incentivo do meu orientador desde o mestrado, propus o projeto de doutorado *Lingüística e Processamento Automático das Línguas Naturais: Explorações Sobre Uma Possível Intersecção* ao Programa de Pós-Graduação em Letras desta unidade da Unesp. Para desenvolvê-lo, de agosto de 1991 a agosto de 1992, contei com o auxílio do Programa da CAPES de Doutorado no País com Estágio Exterior, que tornou possível a realização do imprescindível estágio no exterior, junto ao programa de pós-graduação com o qual estabelecera contato no ano anterior.[1]

Se, de um lado, o estágio me permitiu adquirir conceitos específicos, conhecer a nova dinâmica de pesquisa, estabelecer a necessária visão de conjunto deste campo de estudos de vanguarda e descobrir suas potencialidades acadêmicas e tecnológicas, de outro, o material estudado e a convivência com pesquisadores de diferentes áreas do conhecimento, desenvolvendo os mais variados projetos de sistemas computacionais de processamento automático de línguas naturais, evidenciaram um desconcertante quadro de distanciamento entre estes e os estudos lingüísticos.

Experienciei, mais uma vez, o incômodo distanciamento que separa as "Exatas" das "Humanas". Nesse domínio, ser "apenas" lingüista também não é suficiente! Afinal, o que os lingüistas poderiam entender de processamento automático de línguas naturais?

---

1 Deixo registrada a grande importância do Programa de Doutorado com Estágio no Exterior (**PDEE**), iniciativa louvável e inovadora tomada pela **CAPES**, que a partier deste ano possibilita uma oportunidade ímpar aos pós-graduandos brasileiros de estagiarem junto a centros avançados de pesquisa no exterior, promovendo assim o fortalecimento e o enriquecimento dos programas de pós-graduação, bem como o desenvolvimento, aprimoramento e ampliação de novas frentes de pesquisa no país.

Concluí, assim, que o impecílio não é ser desta ou daquela area, mas ser um especialista em processamento automático de línguas naturais. Mas que especialidade é essa?

Esta tese traduz os esforços que concentrei na árdua tarefa de explorar essa questão, aventurando-me a buscar contornos mais definidos da face tecnológica dos estudos da linguagem que, a meu ver, encontra-se em estado latente no próprio domínio da Teoria Lingüística e difusamente espelhada no vasto e disperso domínio dos estudos sobre o Processamento Automático de Línguas Naturais.

Introdução

> *Conhece os computadores? Indiferentes e tranqüilos, eles tornaram-se as esfinges da nossa civilização moderna, que parece não mais passar sem eles. Encontramo-los por todo o lado: nos jogos, nas fábricas, nos escritórios, nos laboratórios e na televisão.*
>
> Thomas Lachand Robert (1993: 7)

Desde a sua introdução no início dos anos 40, os computadores digitais 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 pesquisas que, sem eles, talvez, nunca teriam sido cogitadas. Destacam-se, por exemplo, a Teoria dos Autômatos, a Teoria das Linguagens Formais, a Teoria dos Algoritmos, a Teoria da Complexidade, a Teorias da Lógica, entre outras (*cf.* KORFHAGE, 1966; TURNER, 1984; BARTON, BERWICK & RISTAD, 1987; SUDKAMP, 1991).

Capazes de proporcionar horas de lazer e entretenimento, de auxiliar na realização de tarefas cotidianas, de resolver, com rapidez e precisão, uma infinidade de problemas complexos e de, até mesmo, substituir em tarefas arriscadas, repetitivas e estafantes, essas esfinges da civilização moderna, admiradas por uns e ignoradas e até temidas muitas vezes por outros, hoje estão indiscutivelmente por toda a parte.

Essas máquinas, que cada vez mais vão fazendo parte de nosso cotidiano e nos auxiliando na construção de conhecimentos

sofisticados, colocaram seus idealizadores diante de um primeiro enigma: como fazê-las decodificar instruções, necessárias para a execução de tarefas?

A criação das *linguagens de programação* foi a resposta imediata que os cientistas encontraram para esse enigma: a comunicação homem-máquina poderia ser estabelecida por meio da "desajeitada" linguagem da máquina.

Para se ter uma idéia mais concreta da dimensão desse problema inicial, basta lembrar que toda a informação armazenada em qualquer computador, mesmo nos computadores de última geração, encontra-se codificada em termos de *bits* e *bytes*.[2] Aos *bits*, forma abreviada do inglês *binary digits*, isto é, dígitos binários, correspondem os dois numerais 0 e 1, utilizados no sistema de representação binária de todos os números. Na verdade, o *bit* pode se entendido como a abstração dos dois estados possíveis – ausência ou presença de corrente elétrica – em que se encontra cada um dos "fios elétricos" que compõem os circuitos, suporte físico de todos os computadores. Por exemplo: o número sessenta e oito, que no sistema decimal é representado pelos algarismos 6 e 8, dispostos na configuração 68, no sistema binário, é representado pelos algarismos 0 e 1, dispostos na configuração 1 0 0 0 1 0 0.[3] Os *bytes*, por sua vez, designam seqüências, em geral, compostas por oito *bits*.

---

2 Na medida do possível, adotarei os termos técnicos de informática propostos para o português de acordo com o ***Glossário de Informática*** (CAMARÃO, 1989), referendado pelo Presidente do Comitê Brasileiro de Informática da ABNT, com as adaptações necessárias em função de nosso objeto específico de estudo.

3 A relação de igualdade entre o número 68, na base decimal, e o número 1000100, na base binária, pode ser explicitada por meio da seguinte fórmula: $\mathbf{6} \times 10^{1} + \mathbf{8} \times 10^{0} = \mathbf{1} \times 2^{6} + \mathbf{0} \times 2^{5} + \mathbf{0} \times 2^{4} + \mathbf{0} \times 2^{3} + \mathbf{1} \times 2^{2} + \mathbf{0} \times 2^{1} + \mathbf{0} \times 2^{0}$.

Tomemos um exemplo concreto. Como mostra a figura a seguir, a palavra DO, por exemplo, estaria armazenada no interior de um computador em dois *bytes*, cada um deles composto de oito *bits*: 0 1 0 0 0 1 0 0, representação binária do número 68, e 0 1 0 0 1 1 1 1, representação binária do número 79. Os números 68 e 79 correspondem aos códigos ASCII das letras D e O, respectivamente.[4]

D							
byte							
0	**1**	**0**	**0**	**0**	**1**	**0**	**0**
bit	bit	bit	bit	bit	bit	bit	bit

O							
byte							
0	**1**	**0**	**0**	**1**	**1**	**1**	**1**
bit	bit	bit	bit	bit	bit	bit	bit

Se o computador, por exemplo, receber a instrução para ordenar alfabeticamente as palavras ODOR e DEDO, ele apresenta, como resposta, a ordem DEDO - ODOR, não porque, no alfabeto, D precede O, mas porque 68 é um número menor que 79. Isso mostra que a única linguagem que os computadores de fato interpretam é a linguagem dos "zeros e uns".

Diante dessa limitação, o homem decidiu, então, adaptar-se à criatura, construindo, assim, uma linguagem que a máquina é capaz de processar, mesmo que isso lhe tenha custado horas e horas de um trabalho insano: codificar linhas e mais linhas de instruções em uma

4 O código ASCII (*American Standard Code for Information Interchange*) foi desenvolvido pelo Instituto de Padrões dos Estados Unidos, em 1968, com a finalidade de padronizar a codificação de todos os caracteres manipuláveis pelo computador. Em todos os computadores, com exceção dos computadores *mainframes* da IBM, que usam o código EBCDIC (*Extended Binary Coded Decimal Interchange Code*), a cada caractere (letra, numeral ou símbolo) e a alguns controles básicos (mover o cursor, marcar o final de um arquivo, suspender a execução de um programa, entre outros), o código ASCII estipula um número específico de identificação interpretável pela máquina (*cf.* CAMARÃO, *op. cit.*: 20 e 171; COVINGTON *et al.*, 1988: 39-40)

linguagem tão misteriosa quanto o próprio enigma. Observe que a seqüência de instruções, a seguir, codificada na linguagem de zeros e uns da máquina, instrui o computador para executar a soma dos números 2 e 4:

0011 1110 0010
1100 0110 0100
0011 0010 0101
0111 0111 1111
1100 1001

A partir de experiências como essa, criaram-se outras linguagens de programação que, aos poucos, foram se distanciando dessa representação imposta pela arquitetura do computador e tornando-se mais inteligíveis, pelo menos do ponto de vista humano.[5] Destaca-se, por exemplo, a linguagem de programação PROLOG.[6] Aquele mesmo conjunto de instruções, agora traduzido para essa linguagem de alto nível, assume a seguinte forma:

Y is 2+4.

---

5 *Cf.* Marshall (1986).

6 A linguagem **PROLOG** (***PRO****gramming* ***LOG****ic*) é uma das mais sofisticadas linguagens de programação para a implementação de programas que envolvem o processamento das línguas naturais. Criada por Alain Colmerauer e seus colegas, na Universidade de Aix-Marseille, em 1972, essa linguagem foi escolhida para o desenvolvimento do *Projeto de Quinta Geração*, projeto japonês, avaliado em um bilhão de dólares, que investiga a criação de computadores capazes de receber instruções codificadas em japonês, (*cf.* COVINGTON *et al.*, *op. cit.*; CLOCKSIN & MELLISH, 1987; TOWNSEND, 1990; ABRAMSON & DAHL, 1989). A linguagem **LISP** (***LIS****t* ***P****rocessing*), outra linguagem de programação também criada para a mesma finalidade, disputa, com a linguagem PROLOG, o lugar de destaque nos projetos de Inteligência Artificial (WINSTON & HORN, 1989).

Uma vez digitada essa instrução, o interpretador PROLOG "responde":

Y = 6

yes

|?–

Para compreender esse "enunciado" cifrado, é preciso saber que o programa que interpreta as instruções em PROLOG, isto é, o *interpretador* PROLOG, primeiro, resolve a operação e, depois, apresenta, no monitor do computador, em linhas consecutivas, as seguintes "frases": **Y = 6**, **yes** e o seu *prompt* característico, formado pela seqüência de símbolos "**|? –**". **Y = 6** expressa o resultado da operação solicitada, a palavra inglesa **yes** assevera que a operação foi resolvida com sucesso e a seqüência de símbolos **|? –**, um "marcador fático", sinaliza para o usuário que o canal de comunicação continua aberto, à espera de novas instruções.

Embora a instrução codificada em PROLOG seja indiscutivelmente muito mais inteligível que as seqüências enigmáticas da linguagem da máquina, ela evidentemente não é uma instrução codificada em inglês. Se não digitarmos a instrução exatamente da forma prescrita pela linguagem PROLOG, isto é, **Y is 2 + 4.**, com a variável **Y** escrita em maiúscula, a seqüência **is** com letras minúsculas e o característico ponto final, receberemos – frustrados – um **no** (não) ou um **syntax error** (erro sintático) como resposta.

Cientes dessa inevitável rigidez, muitos pesquisadores se propuseram a pensar sobre possibilidades de fazer com que os computadores se transformassem em instrumentos mais acessíveis. Uma

das saídas encontradas foi a construção de interfaces gráficas, isto é, programas que transformam a informação em objetos gráficos e que servem de veículo de comunicação entre o usuário e o computador. A questão colocada foi: por que não criar "máscaras" que escondam essa maneira primitiva de comunicação? Essa alternativa, hoje, parece ter sido resolvida com grande sucesso. Os computadores modernos, de fato, dispõem de sofisticadas "máscaras". A "linguagem das interfaces gráficas", com seus menus, ícones e cores, não só oculta o que realmente se passa dentro de um computador, mas também os transforma em máquinas muito mais atraentes e fáceis de operar, uma vez que o usuário não precisa mais digitar dezenas de comandos muitas vezes obscuros e de difícil memorização.[7]

Uma outra possibilidade, cuja realização é sem dúvida muito mais complexa, continua sendo um desafio: criar programas de computador capazes de interpretar mensagens codificadas em línguas naturais. Por que não investigar meios que façam com que as máquinas "aprendam" nossa própria linguagem e sejam capazes de decifrá-la e usá-la?

Com efeito, essa preocupação com a comunicação "mais natural" entre o homem e a máquina já se instalava, desde o momento da própria criação dos primeiros computadores (PYLYSHYN, 1980: 463):

> *"Ever since the early days of computing, researchers have been intrigued by the idea of communicating easily with computers. In*

---

[7] As interfaces gráficas, ou "plataformas gráficas", a que me refiro, começaram a ser desenvolvidas a partir da década de 80. Em 1984, a empresa norte-americana Apple, com o apoio da empresa Xerox, colocava no mercado o *Macintosh*™, o primeiro computador pessoal equipado com uma sofisticada interface gráfica. Depois dessa iniciativa, a Microsoft e a IBM, outras empresas norte-americanas, os gigantes da informática, também passaram a desenvolver suas próprias plataformas gráficas: a série *Windows*™ e o sistema operacional *OS2*™, respectivamente.

> *nearly every area of computing, one can imagine how the understanding of language could make computers more accessible, not only for those who use them but for many laymen."*

As preocupações, porém, foram muito mais além. Por que não ousar? Por que não criar meios que instruam o computador a transformar, por exemplo, a citação acima em:

> **"Desde os primórdios da computação, os pesquisadores são fascinados pela idéia de se comunicarem facilmente com os computadores. Em quase todas as áreas da computação, é possível imaginar como a compreensão das linguas poderia tornar os computadores mais acessíveis, não só para aqueles que os usam, mas também para muitos leigos" ?**

Questões como essas evocam o grande enigma que as esfinges deste século XX reservavam àqueles que iriam se aventurar a decifrá-las: como fazê-las "compreender" a linguagem humana?

Posto o grande enigma, inúmeros "aventureiros" se dispuseram a criar meios para decifrá-lo. Desde então, criar programas computacionais "inteligentes" e capazes de "compreender" as línguas e, por meio delas, simular uma interação verbal com o usuário, tem se revelado um empreendimento polêmico, complexo e desafiador, porém, extremamente fascinante.

Nesta tese, passo a empregar o termo "processamento automático de línguas naturais" (PLN) para denotar especificamente o objeto da pesquisa:[8] desenvolvimento de sistemas computacionais capazes de processar objetos de natureza lingüística.

A grande meta prevista para as pesquisas dessa natureza é indiscutivelmente ousada: projetar e implementar sistemas

[8] Por extensão, o termo PLN será também empregado para denotar o campo de estudos delineado neste trabalho.

computacionais avançados em que a comunicação entre o homem e o computador possa realizar-se por meio de línguas naturais, e não por meio de instruções e comandos codificados numa linguagem de programação artificialmente construída por programadores. Assim, investigar o PLN é, antes de tudo, aventurar-se em participar de um empreendimento fascinante e desafiador que, talvez um dia, venha a transformar máquinas em nossos "interlocutores e parceiros cibernéticos", capazes de nos auxiliar no planejamento das mais variadas tarefas e, até mesmo, na resolução dos problemas mais recalcitrantes.

Hoje, com quase meio século de experiências acumuladas nesse sentido, algumas bem-sucedidas, outras absolutamente desastrosas, o PLN apresenta-se como um campo de estudos bastante heterogêneo e fragmentado, acumulando uma vasta literatura e agregando pesquisadores das mais variadas especialidades, com formação acadêmica, embasamento teórico e interesses também bastante diversos. O mais agravante, porém, é constatar que os complexos fenômenos da linguagem, cuja compreensão é condição essencial para o sucesso do empreendimento, têm sido, muitas vezes, ingênua ou descuidadamente subdimensionados, evidência de um incômodo e pernicioso distanciamento entre os estudos do PLN e os Estudos da Linguísticos.

Diante desse quadro caótico e desnorteador, arrisco colocar parte das experiências bem-sucedidadas em perspectiva, na tentativa de propor uma caracterização integrada do PLN que possibilite e estimule a realização de trabalho solidário. Com isso, espero contribuir para delinear uma face tecnológica para os estudos da linguagem e minimizar a lacuna que separa esses dois domínios, divulgando e incentivando esse

modo de investigação científica e tecnológica no âmbito das Humanidades.

Para atingir esses objetivos, aponto uma série de problemas que considero conjunturais e, a partir dessa reflexão, proponho uma estratégia de pesquisa e um enfoque do PLN que viabilize e estimule o trabalho cooperativo entre as equipes de especialistas.

Em termos formais, esta tese organiza-se, além desta introdução e das "Conclusões e Perspectivas", em seis capítulos. No primeiro, aponto os problemas conjunturais que considero entraves para o trabalho cooperativo entre lingüistas e projetistas de sistemas de PLN. No segundo, proponho o equacionamento estratégico global para o empreendimento, evidenciando a importância de se construir sistemas de PLN lingüisticamente motivados, delimitando a concepção de PLN e salientando o papel dos estudos do PLN enquanto gerador de pesquisas acadêmicas e tecnológicas. No terceiro, explicito, de modo sistemático, as relações de interdisciplinaridade que se estabelecem entre as pesquisas do PLN e as disciplinas matrizes que lhe dão fundamentação, sistematizo os recursos teóricos para o desenvolvimento dos projetos e apresento a estratégia de pesquisa integrada para o PLN que busca o equacionamento dos problemas em três domínios: o Lingüístico, o Representacional (Linguístico-computacional) e o Implementacional (Computacional). Como decorrência, proponho também uma estratégia de pesquisa para a construção de um sistema de PLN particular. Nos três capítulos subseqüentes, equaciono os principais problemas nos três domínios: o Lingüístico (quarto capítulo), o Representacional (quinto capítulo) e o Implementacional (sexto capítulo).

CAPÍTULO 1 – Os desafios

> *‘‘The fragmentation of the field of linguistics and the fuzzy philosophizing that passes for ‘linguistic theory’ among large segments of the linguistic population don't inspire much confidence among the language engineers, and the blissful ignorance about elementary facts of natural language that the engineers flaunt smugly in their publications does little to convince serious linguists that there is anybody out there among mainframes who has any interest in applying whatever linguistic scientists may have found out…”*
>
> Stanley Starosta (1991: 178)

Construir um corpo de conhecimentos suficientemente estruturados e integrados, capaz de fornecer os meios que poderão tranformar máquinas em “tradutores ou interlocutores cibernéticos” é um empreendimento arrojado e fascinante. Entretanto, ao me aventurar em participar desse empreendimento, constatei o fato intrigante descrito na epígrafe deste capítulo: a existência de uma preocupante lacuna entre os Estudos Linguísticos e do PLN.

Desafios para os projetistas

Do ponto de vista dos projetistas de sistemas de PLN, é possível encontrar razões que os levam a se distanciar dos Estudos Lingüísticos. Como ponto de partida, relembro parte das dificuldades

que enfrentei ao desenvolver o meu estudo do Mestrado (DIAS-DA-SILVA, 1990) no domínio da Teoria Lingüística. Na busca do melhor caminho que me levasse a uma compreensão maior do fenômeno da apassivação nas línguas naturais, tema bastante recorrente nesse domínio, andei às voltas com um "frenético borbulhamento de novidades teóricas". Escolher um arcabouço teórico 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".

Essas mesmas expressões, tomadas de empréstimo de Lemle (1984: 2), ainda traduzem parte dos problemas com os quais nos deparamos diante da necessidade de escolhas no âmbito da Teoria Lingüística. É também oportuno esclarecer que Lemle fez uma escolha teórica ao realizar trabalho semelhante: propor uma "ponte entre a lingüística teórica e o ensino escolar da gramática". Revelador também é o fato da lingüista atacar os seus pares, ao dizer que "uma certa lingüística" emprega um "formalismo algébrico bizarro, abominável e desinteressante para a maioria das pessoas" e ao criticar severamente a Teoria Lingüística, afirmando que o discurso lingüístico vale-se com "demasiada freqüência de um linguajar técnico hermético que disfarça o vazio de substância de suas propostas".

Essa experiência revela que escolher e avaliar esquemas teóricos no âmbito da lingüística transformam-se em processos complexos, laboriosos e, principalmente, desnorteadores. Não raro, as propostas apresentadas pelos teóricos trazem consigo uma pluralidade de análises, muitas delas simplesmente esboçadas, uma metalinguagem, de

fato, hermética, propositalmente cifrada, e uma evidente concentração em aspectos particulares e pontuais da análise das línguas. Qualquer lingüista reconhece o viés sintático que dominou, e de certa forma ainda domina, as pesquisas lingüísticas.

A fragmentação, a parcialidade e a pouca formalização dos modelos lingüísticos são também apontadas como agravantes para o quadro de distanciamento. Winograd (1972: 41), por exemplo, justifica-se:

> "Quando começaram os primeiros trabalhos de análise das línguas naturais por meio de computadores, não havia teorias sintáticas suficientemente explícitas, prontas para receberem um tratamento computacional. Os projetistas pioneiros que ousaram propor os primeiros sistemas de tradução automática foram forçados a construir seus próprios modelos lingüísticos, à medida que seus projetos desenvolviam-se. Como decorrência, eram modelos extremamente 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: 1), comentando que os lingüistas tratavam-na como algo "excelente para se discutir, porém incognoscível", possuindo as "mesmas qualidades de Deus ou da Mente", diz:

> "Uma vez 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 & Riesbeck (1981: 2) acusam os lingüistas não só de se fixarem demasiadamente nos estudos sintáticos e minimizarem os estudos semânticos mas, sobretudo, de pouco se preocuparem com o estudo do desempenho lingüístico:

> "Quando os primeiros trabalhos sobre PLN começaram, os pesquisadores passaram a experimentar qualquer teoria disponível. As implementações de inúmeras teorias lingüísticas, enfatizando a Sintaxe, tiveram uma certa popularidade no âmbito da Inteligência Artificial durante algum tempo. Mas a verdadeira procupação sempre foi o tratamento do significado, ao passo que, durante um longo tempo, os lingüistas evitaram ao máximo abordar essa questão. Quando os lingüistas decidiram, de fato, abordar as questões do significado, não o fizeram do ponto de vista do processo… Os pesquisadores engajados com o PLN, no âmbito da Inteligência Artificial, tiveram de enfrentar a tarefa sozinhos, propondo suas próprias teorias do processamento lingüístico."

Ainda hoje, a crítica aos lingüistas continua ecoando (MYKOWIECKA, 1991: 497):

> "Um dos motivos que vêm impedindo o rápido desenvolvimento do campo 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 da Informática: cada grupo trabalha para oferecer ao mercado consumidor programas mais sofisticados, mais rápidos e mais econômicos que os

programas desenvolvidos por seus pares. Indignado, esse pesquisador acrescenta que, por esse motivo, é muito difícil encontrar publicações que informem, com regularidade, os avanços alcançados nos laboratórios 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 por aplicativos e as limitações de recursos computacionais, por exemplo, têm impedido que os sistemas de PLN passem a incorporar sofisticações que exijam estudos mais aprofundados e consistentes com as descobertas da Teoria Lingüística. Há justificativas plausíveis do ponto de vista comercial e mercadológico, mas que não se sustentam do ponto vista acadêmico e tecnológico. Allen (1987: 2) já alertava para esse risco. Ele é categórico, ao afirmar que os objetivos tecnológicos não poderão ser alcançados sem a busca de fundamentação nas sofisticadas teorias propostas no âmbito da Lingüística Teórica.

A pseudo-autonomia dos Estudos do PLN em relação aos Estudos da Linguagem acaba sendo justificada, de fato, se considerarmos que, de certo modo, a Lingüística não tem se preocupado em auxiliar os trabalhos de PLN. Sempre ocupados com o estudo da linguagem humana *per se*, os lingüistas ficam circunscritos aos limites de sua atuação. Intencionalmente ou não, deixam transparecer um certo descaso, resistindo a cooperar com projetos voltados para o PLN e, principalmente, ignorando a importância crucial da sua contribuição para o avanço desse corpo de conhecimentos interdisciplinares. O lingüista Halvorsen (1989), mesmo defendendo a necessidade de maior interação entre as duas áreas, chega a comentar que a Teoria Lingüística, apesar de

reunir dados significativos sobre o complexo desempenho lingüístico humano, ao invés de incentivar, tem muitas vezes desestimulado as pesquisas sobre o PLN.

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 Lingüistica. O emprego dos termos 'discurso' e 'texto' constitui 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 e têm gerado, como se mostrarão nos parágrafos seguintes, debates e embates entre teóricos.[9]

Dubois *et al.* (1978: 192) apresentam três noções distintas de discurso: (i) "linguagem posta em ação", (ii) "uma unidade igual ou superior à frase" e (iii) "todo enunciado superior à frase". Explicam que "na problemática anterior à análise do discurso […] a oposição enunciado/discurso marcava simplesmente a oposição entre lingüístico e não-lingüístico […] O estudo dos processos discursivos que justificam o encadeamento das seqüências de frases eram remetidos à psicanálise […] Benveniste [é quem] propõe como lingüístico o problema do discurso." Já, para esses autores, o texto é tomado ora como discurso ora como *corpus* de enunciados lingüísticos.

---

9 Para uma apreciação detalhada das dificuldades e confusões causadas pela pluralidade de usos desses termos, remeto o leitor para Fávero & Koch (1983: 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 & Courtés (1979) e Heydrich, *et al.* (1989). 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 a ''Lingüística da Frase'' e a ''Lingüística Transfrástica'': *análise do discurso*, *lingüística textual*, *gramática* ou *sintaxe do texto*, *análise da conversação*.

Hatim & Mason (1990: 243) propõem uma conceituação funcional para texto e atitudinal para discurso. Para esses autores, o texto define-se pela intenção global do seu criador, instanciada pelo "conjunto de funções comunicativas mutuamente relevantes, estruturadas de forma a atingir um propósito retórico global" como, por exemplo, a intenção de narrar ou contra-argumentar. Como critérios de textualidade, destacam dois tipos de relações: as de coesão e as de coerência. Já o discurso caracteriza-se pela atitude que os participantes adotam em relação, por exemplo, a áreas de atividades socio-cultural, caracterizando, assim, o "discurso racist" e "discurso ufanista", entre outros.

Velde (1989: 175) emprega os termos 'texto' e 'discurso' indistintamente para denotar uma "seqüência/conjunto de enunciados verbais reconhecível/identificável como um todo coerente".

Lyons (1977: 30) define texto em função de discurso, não atribuindo estatuto teórico algum para este último: "Empregaremos o termo *texto* para designar qualquer passagem conexa de discurso, quer ela seja falada ou escrita, quer ela seja um diálogo ou um monólogo.[10]

Enkvist (1989: 370-2), aparentemente discordando dos demais autores, sugere uma distinção e propõe um inter-relacionamento entre os dois termos: o texto é caracterizado como "uma seqüência significativa de símbolos em uma língua natural", e o discurso como a soma do texto mais o seu contexto situacional. Acrescenta, porém, que "discurso" e "texto" poderiam ser tomados como sinônimos, se admitirmos que texto e contexto situacional são entidades inseparáveis.

---

10 Para Lyons, portanto, o discurso é sinônimo de *fala*, isto é, a "língua em uso".

Nos trabalhos de PLN, o discurso é, em geral, concebido como "qualquer segmento conexo de texto ou fala, compreendendo uma ou mais frases ou segmento de frases" (SIDNER, 1979: 122). Nessa definição, é possível inferir, a partir da oposição texto/fala, que 'texto' denota um "discurso escrito". Conceituação semelhante é também apresentada por Leech (1983: 59), para quem discurso é um ato de fala e texto é uma realidade física, uma seqüência de sons ou símbolos gráficos. Ou, como tenta esclarecer Grishman (1986: 141),

> "Empregaremos o termo **discurso** para designar qualquer texto que contenha mais de uma frase. O discurso pode concretizar-se em uma multiplicidade de formas. O discurso pode narrar uma história, descrever uma cena, fornecer instruções ou encerrar um argumento."

Ou ainda, como preferem Scha *et al.* (1990: 233),

> "O termo **compreensão do discurso** refere-se a todos os processos de compreensão das línguas naturais que visam à interpretação de um texto ou diálogo. Para tais processos, cada frase da língua natural é um elemento cuja importância reside na sua contribuição para o engendramento de significados de segmentos maiores, e não no seu significado individual. Para compreender o discurso, é preciso mapear a estrutura do texto, ou diálogo, à medida que este vai se desdobrando, e interpretar cada enunciado subseqüente em função do contexto apropriado – levando-se em consideração tanto o contexto situacional dos enunciados quanto o co-texto lingüístico formado pelos enunciados precedentes."

O emprego desses termos como conceitos meramente operatórios é claramente revelado por Grosz & Sidner (1986: 176), ao proporem o seu modelo computacional das Estruturas do Discurso:

> "Embora tenhamos de pospor a proposição de uma definição para discurso até que a teoria aqui apresentada contenha elementos que justifiquem essa síntese, algumas propriedades dos fenômenos a ele relacionadas, e que pretendemos explicar, podem desde já serem especificadas. Por ora, identificaremos o discurso com uma parcela de comportamento verbal, envolvendo tipicamente enunciados e participantes múltiplos."

Os empregos discrepantes dos termos 'discurso' e 'texto' parecem, porém, esconder problemas muito mais sérios. A que objeto, ou objetos, esses termos se aplicam? Discurso e texto são objetos distintos? São faces distintas de um mesmo objeto? São dimensões distintas de um mesmo objeto? São simplesmente estipulações? São processos? São produtos? Definitivamente, não há respostas categóricas para essas questões. Infelizmente, a adoção de uma ou outra acepção para esses termos está longe de ser pacífica. Enkvist (*op. cit.*), por exemplo, não só reitera a necessidade de delimitação desses dois objetos mas, principalmente, amplia a discussão, nela incluindo a necessidade de especificação de outros tantos termos – *gramaticalidade*, *aceitabilidade*, *propriedade*, *encaixe textual*, *saliência*, *coesão*, *coerência*, *interpretabilidade* e *conectividade* – não menos controvertidos.

Frente à variedade de posturas diante desses e de tantos outros fenômenos lingüísticos, não é de causar surpresa que os próprios lingüistas e, sobretudo, os projetistas de sistemas de PLN sintam-se desnorteados e acabem por adotar modelos díspares ou, até mesmo, por

criar seus próprios modelos, muitas vezes *ad hoc*, aumentando assim os desencontros.

Desafios para os lingüistas

Se os projetistas de PLN podem valer-se de uma série de argumentos para continuarem se distanciando dos lingüistas, estes também encontrarão argumentos de sobra para não se engajarem em projetos sobre o PLN.

O domínio do PLN agrega uma heterogeneidade de objetivos. Encontram-se projetos voltados para a utilização do computador como uma simples ferramenta auxiliar da pesquisa principal como, por exemplo, o uso de programas que calculam estatísticas de ocorrências de palavras em textos em geral ou que possibilitam a indexação de palavras e segmentos de textos, até projetos extremamente ambiciosos que estabelecem como meta a criação de uma "inteligência artificial" nos moldes do super-computador HAL, dotado, entre outras, da capacidade humana da linguagem, personagem central do filme clássico de Stanley Kubrick — *2001:Uma Odisséia no Espaço*.

Em outras palavras, há uma considerável pluralidade de objetivos: desde o estudo quantitativo das línguas naturais, que, na essência, visa à construção de listas de freqüência de palavras e à análise de possibilidades combinatórias de unidades lingüísticas, passando pelo estudo da adequação formal e psicológica de modelos de descrição lingüística, por meio da implementação computacional de gramáticas, até a proposição de sofisticados modelos computacionais que "dialogam" com o usuário ou que são capazes de "compreender histórias".

Nesse meio, há também uma série de sistemas de PLN muito rudimentares e, em geral, desprovidos de qualquer embasamento lingüístico. Basta citar, 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. A "tecnologia lingüística" nesses aplicativos é praticamente inexistente. Todos os elementos lingüísticos envolvidos são manipulados segundo técnicas de indexação e de algoritmos que contornam os problemas computacionais postos pela complexidade das línguas naturais.

Os "tradutores de bolso" são outro exemplo de aplicativo que não apresenta qualquer vestígio de PLN. Esses dispositivos do tamanho das calculadoras convencionais são simplesmente equipados com listas de palavras e expressões de línguas diversas, algumas frases e fragmentos de frases. O programa subjacente limita-se a manipular esses elementos: por meio de comparações, detecta as palavras equivalentes de línguas diferentes e, por meio de algumas 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 sintática.

O fato é que parcelas muito pequenas dos resultados de pesquisas pioneiras sobre o PLN têm sido timidamente incorporadas em uma variedade de produtos: determinados *games* de computador, que aparentemente comunicam-se usando fragmentos de línguas naturais; os pequenos dicionários e tradutores eletrônicos de bolso, que fornecem a tradução de palavras, expressões e frases em diversas línguas; os

diversos processadores de texto, equipados com “corretores ortográficos”, “dicionários de sinônimos e antônimos”, entre outros aplicativos; os dicionários e as enciclopédias informatizados, que podem ser consultados *on-line*; e os sistemas informatizados de acesso a base de dados por meio de perguntas em uma pseudo-linguagem natural.

Todavia, no presente estágio de desenvolvimento dessas tecnologias, o que observamos são implementações de fórmulas lingüísticas estereotipadas. O mesmo acontece com os “sistemas especializados” comercializados, que fornecem informações sobre um determinado tópico em forma de textos previamente armazenados no sistema.

Os estudos sobre o PLN, até por isso, são muitas vezes rotulados de ecléticos, consumistas, imediatistas e puramente comerciais, interessados apenas em “consumir’’ quaisquer contribuições de outras áreas do conhecimento que lhes sejam úteis – Filosofia, Lógica, Psicologia, Lingüística, Ciência da Computação e Inteligência Artificial.

Do ponto de vista teórico-metodológico, avaliar propostas e selecionar estratégias de trabalho transformam-se em problemas ainda mais complexos. Por ser um campo incipiente e heterogêneo, que vem sendo explorado por pesquisadores de áreas bastante diversas, deparamo-nos com uma variedade de propostas, ferramentas e equipamentos computacionais, cuja sistematização torna-se praticamente impossível. Grande parte desses resultados 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.[11]

Nessa efervescência, encontram-se análises estatísticas, sistemas lógicos, teoria dos grafos, teoria dos conjuntos, teoria de modelos, teoria das linguagens formais, teoria dos algoritmos, teoria da complexidade, representação do conhecimento, entre outras. Evidentemente, parte dessas teorias é também utilizada por lingüistas como, por exemplo, a teoria dos grafos e a teoria das linguagens formais, que, a partir da proposição da gramática gerativo-transformacional (CHOMSKY, 1957), são amplamente empregadas como esquemas de descrição e representação das regras e estruturas sintáticas das línguas. Vejo, até mesmo, um lado positivo nessa diversidade. Do ponto de vista de recursos formais, imprescindíveis para o tratamento computacional das línguas, há que se reconhecer, contudo, que o conjunto desses trabalhos constitui um referencial rico, sugerindo modelos e sofisticadas técnicas de representação e manipulação do material lingüístico.

O levantamento de projetos e de aplicativos revela que o PLN está imerso em um domínio de pesquisa difuso, controvertido e caótico. Um domínio à espera de organização que assinale contornos mais claros e identifique seus objetos, criando 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 se admitindo que não há um referencial único, que modelos explícitos e completos ainda estão para serem construídos e que "lutas teóricas" fazem parte de sua história (LEECH, *op. cit.*), há que se concordar que existem parâmetros

11 Obras como Grishman (*op. cit.*), Allen (*op. cit.*) e Gazdar & Mellish (1989) estão entre as poucas tentativas de reunir didaticamente os temas pertinentes sobre o PLN.

norteadores mínimos a que os lingüistas, mesmo diante dos problemas apontados, recorrem para ancorar as suas investigaçõe. Verificamos certo consenso 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 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, entre outros (*cf.* JAKOBSON, 1977; BORBA, 1984; SELLS, 1985).

Essa base comum, resultante de uma longa tradição de estudos gramaticais, acaba por fornecer um universo de discurso comum, contendo uma metalinguagem e noções gerais, fato que se evidencia no discurso dos próprios teóricos da linguagem, que constantemente recorrem a ela para construir suas análises.[12] Uma breve leitura de 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 novos conceitos. Para exemplificar, destaco este trecho inicial da discussão sobre a natureza das relações gramaticais (MARANTZ, 1984: 1):

---

12 Fato também revelador de desencontros é ter de admitir que a Gramática Tradicional, espelhada nos vários manuais, fora dos círculos lingüísticos, ainda continua sendo a principal fonte de referência sobre as línguas vernáculas e estrangeiras.

> "Os lingüistas têm razoável clareza sobre noções gerais, noções pré-teóricas como 'antecedente de um pronome reflexivo', 'especificação de casos morfológicos', 'ordem das palavras', 'agente de uma ação', e outros conceitos que parecem estar relacionados com a noção de 'sujeito'. Mas o 'sujeito' propriamente dito não pertence a essa classe de conceitos pré-teóricos. Como, então, é possível avaliar uma proposta teórica das relações gramaticais ou a definição de sujeito e objeto? "**13**

Envolver-se com o PLN, porém, implica estar disposto a compreender uma efervescência de teorias e técnicas emprestadas das mais variadas areas e a decifrar formalismos algébricos muito mais bizarros que os mencionados por Lemle.

Em se tratando de um campo de pesquisa em que parcelas de conhecimentos precisam (e devem) ser cuidadosamente extraídas de domínios de estudos diversos, os estudos do PLN reservam uma outra dificuldade: apresentar uma metalinguagem fragmentada, um hermetismo terminológico e, até mesmo, uma desnecessária multiplicidade de termos exóticos.**14** A terminologia que Shapiro (1990) discute para descrever os tipos de processamento ilustra a proliferação de termos exóticos, provenientes das áreas da Ciência da Computação e da Inteligência Artificial. Os três pares de modificadores *bottom-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

13 Grifo meu.

14 Observo que essa metalinguagem exótica mereceria um estudo *per se*, que poderia também contribuir para a minimização de desencontros.

modificar os termos *chaining* (encadeamento), *inference* (inferência), *parsing* (análise sintática), *processing* (processamento), *reasoning* (raciocínio) e *search* (busca, pesquisa), gerando termos como: *bottom-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 termos novos, mais precisos e específicos. A sua proliferação assistemática, porém, acaba por criar dificuldades adicionais que precisam ser contornadas.

Lyons (1977), nos anos 70, assinalava a necessidade de enfrentar tarefa semelhante no âmbito dos estudos sobre a 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, chega a reconhecer que o tratamento terminológico e conceitual transformou-se em uma de suas maiores dificuldades. Alertava, então, para o perigo de duas situações: (i) autores diversos atribuírem acepções distintas a um mesmo termo, e (ii) autores diversos empregarem termos distintos para descrever fenômenos essencialmente idênticos. Após tentativas frustradas de encontrar uma maneira de compatibilizar as discrepâncias, acabou optando por apontá-las apenas 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, são freqüentemente transportados, sem constrangimento algum, para o universo dos computadores. Nesse universo humanóide, não é difícil encontrar máquinas que pensam, máquinas consultoras, prontas para estabelecer diagnósticos, fornecer consultoria e dar conselhos, máquinas que conversam não só entre si como também com usuários humanos, máquinas poliglotas, máquinas tutoras, planejadoras e até máquinas aprendizes.

Minsky (1968: 2), chega a defender até mesmo 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."

Observemos estes dois exemplos adaptados de Gazdar & Mellish (*op. cit.*: 153).

No primeiro exemplo, a máquina fica "conjecturando introspectivamente", enquanto aplica um possível algoritmo de análise sintática à frase *A casa caiu*.

Eu estou procurando uma frase.
De que é composta uma frase?
Uma frase pode ser composta de um SN seguido de um SV.
Logo, primeiro preciso procurar o SN.
De que é composto um SN?
Um SN pode ser composto de um DET seguido de um N.
Logo, primeiro preciso procurar o DET.

Há uma entrada lexical "a" da categoria DET.
A primeira palavra da cadeia de palavras é "a"?
Sim.
Então encontrei o DET: a palavra "a".
Agora preciso procurar o N.
Há uma entrada lexical "casa" da categoria N.
A segunda palavra da cadeia de palavras é "casa"?
Sim.
Então encontrei o N: a palavra "casa".
Então consegui encontrar o SN: a seqüência "a casa".
Agora preciso procurar o SV.
De que é composto um SV?
Um SV pode ser composto de um V.
Logo, preciso procurar o V.
Há uma entrada lexical "caiu" da categoria V.
A primeira palavra da cadeia de palavras é "caiu"?
Sim.
Então encontrei o V: a palavra "caiu".
Então consegui encontrar o SV: a palavra "caiu".
Então consegui encontrar a frase: a seqüência “a casa caiu”.

No segundo exemplo, os autores comentam as deficiências de uma estratégia de especificação computacional da estrutura sintática de frases, conhecida como “*bottom-up parsing*” (“análise sintática ascendente”)

> “O ingênuo analisador sintático da Seção 5.2 nunca formulava hipóteses sobre o que ele estava procurando, ou delas fazia uso, para decidir sobre seu próximo passo. Ele apenas verificava regras para ver se havia uma maneira lícita de combinar as partes de que dispunha naquele momento. É por isso que ele se via às voltas com as regras que envolviam o agrupamento de ‘espaços em branco’ ”.

A antropomorfização da máquina pode ser justificável do ponto de vista da inteligibilidade das explicações. Essa estratégia discursiva, porém, acaba gerando 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 de termos precisos para descrever os seus objetos.

A precária troca de trabalhos sobre o PLN entre linguistas brasileiros e 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, ao lado de termos bizarros, apresentam também problemas de precisão e até confusões conceituais.

A própria denominação "processamento automático de línguas naturais", expressão que venho empregando como equivalente à expressão inglesa "*automatic natural language processing*", cunhada pelos estudiosos do campo da Inteligência Artificial e da Ciência da Computação, exige alguns esclarecimentos, uma vez que a denominação mais corrente em português parece ser "processamento da linguagem natural". Além disso, o termo inglês "*natural language processing*" e o termo correspondente em português, embora sejam mais freqüentemente empregados para denotar o PLN, são expressões ambígüas, pois podem também denotar o estudo do processamento humano das línguas naturais, objeto de estudo específico da Ciência Cognitiva.

O termo "processamento automático" parece não ser motivo de controvérsias. No sentido usual, denota a utilização de computadores para a estruturação e manipulação de símbolos em geral em que a intervenção humana é reduzida ao mínimo. Essas operações são executadas segundo representações precisas e explícitas, implementadas por meio de programas específicos, escritos em alguma linguagem de programação apropriada.

Já a adoção do termo 'linguagem', de fato, requer esclarecimentos, porque os pesquisadores da área das "Exatas", acostumados a trabalhar com as linguagens formais, muitas vezes,

acabam por empregá-lo de maneira inadequada, gerando confusões desnecessárias e, principalmente, revelando desconhecimento de conceitos-chaves da Teoria Lingüística.

Exemplifico o problema, citando segmentos da tradução para o português de um livro clássico sobre Inteligência Artificial, escrito por uma pesquisadora norte-americana. No capítulo "Compreensão de Linguagem (*sic*) Natural" (RICH, 1983: 344-406), encontrei, desde o título, o termo 'linguagem' que, apropriadamente, deveria ter sido substituído pelo 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."**15**

Nesses contextos, fica evidente que a autora refere-se a línguas e não à capacidade humana da linguagem.

Contribui para a tradução equivocada, o substantivo "*language*" do inglês, que é notadamente ambígüo: ora é empregado para denotar a linguagem, uma das faculdades cognitivas humanas universais, ora é empregado para denotar uma língua em particular, ou seja, uma realização específica dessa faculdade.

Lyons (1981: 16) esclarece a questão:

---

15 Grifo meu.

> "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']."

Akmajian *et al.* (1986: 6) comentam a estranheza manifestada por alunos norte-americanos, quando descobrem que o termo inglês *'language'* possui também o sentido de "faculdade da linguagem" e acrescentam que, para os norte-americanos, a distinção entre língua e linguagem parece ser praticamente desconhecida fora do círculo dos linguistas.

Assim, essas considerações reforçam o meu cuidado com a precisão terminológico-conceitual do campo do PLN. No português acadêmico, o problema do emprego dos termos 'língua' e 'linguagem' nem mesmo se coloca, uma vez que cada um deles reveste conceitos distintos.

Note-se que o termo 'linguagem natural', segundo a tradição gramatical, deve ser reservado para designar genericamente a linguagem humana. Como já se disse, trata-se de uma aptidão característica da espécie humana, cuja manifestação se dá no conjunto das línguas naturais. Não fossem o esperanto (considerado por lingüistas um exemplo inequívoco de língua artificial, talvez por ter como substrato as línguas naturais pré-existentes) e a possibilidade de criação

de outras línguas similares, a qualificação de "natural" seria absolutamente desnecessária.

Já 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, e não naturais.[16]

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

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

---

16 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.

Os termos "casamento" e "mapeamento" são aqui 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). O termo "*matching*", por sua vez, refere-se a um dos processos mais estudados no âmbito da Inteligência Artificial: "*pattern matching*" (SLAGLE & GINI, 1990). Por razões que ficarão explicitadas no quinto capítulo, por ocasião da apresentação da Teoria Léxico-Funcional, proposta pela lingüista Joan Bresnan no início da década de 80, os termos "configuração", "projeção" e "unificação" parecem ser as traduções mais apropriadas para os termos "*pattern*", "*mapping*" e "*matching*", respectivamente.[17]

---

17 O processo de "*pattern matching*" ("unificação de estruturas"), também conhecido como "*unification*" ("unificação") ou "*merging*" ("união"), é também de importância capital para a teoria léxico-funcional.

Desafios para ambos

Além dessas razões conceituais e terminológicas, grande parte dos ataques recíprocos entre os pesquisadores das duas areas parece ser decorrência de um outro fato, também amplamente percebido na academia: a imagem estereotipada e até mesmo distorcida que os pesquisadores formam uns dos outros, sobretudo, se trabalham em domínios de conhecimento diversos.[18] Não é difícil constatar que uma simples conversa entre colegas de áreas distintas é, não raro, pontuada por esses julgamentos provocativos.

John Lyons, por exemplo, no *Prefácio* de uma de suas obras clássicas, *Introdução à Lingüística Teórica* (LYONS, 1979), deixa transparecer os dois esteriótipos, também "clássicos", que há muito têm sido atribuídos a pesquisadores das duas áreas. Num momento em que modelos formais de gramática passavam a ser o centro das investigações lingüísticas, Lyons advertia seus leitores, em especial aqueles cuja formação intelectual se apoiava mais nas ‘‘Humanidades’’, para o fato de que eles deveriam estar preparados para fazer "um certo esforço intelectual com respeito ao uso de símbolos e de fórmulas.’’[19]

Além do rótulo explícito – ‘‘colegas dos números’’, é possível ler nas entrelinhas que os pesquisadores das ‘‘ciências’’ são fatalmente caracterizados como indivíduos pouco intuitivos no que se

---

18 Esse fato talvez seja apenas um reflexo do modelo compartimentado de pesquisa que gerou um modo individualista de pesquisar, atitude que acaba por construir barreiras entre as áreas do conhecimento, sendo, muitas vezes, responsável pela criação de estereótipos que contribuem ainda mais para o distanciamento entre os pesquisadores.

19 Hoje, passados mais de 25 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 são lugar-comum em qualquer investigação 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.

refere à “apreciação dos vários matizes da língua”, academicamente mal formados, uma vez que parecem desconhecer os fundamentos históricos e filosóficos dos Estudos da Linguagem e responsáveis pela criação e proliferação de uma desencorajadora quantidade de ‘‘símbolos e fórmulas arcanas’’, cuja compreensão exige alta capacidade cognitiva. Já os pesquisadores das ‘‘humanidades’’ são tipicamente caracterizados como aqueles de ‘‘hábitos mentais mais voltados para as Letras’’, mais capazes de fazer uma ‘‘apreciação intuitiva dos vários matizes da língua’’, em geral, conhecedores dos fundamentos históricos e filosóficos, 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 até hoje, 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.

É preciso esclarecer, porém, que Lyons combate essa visão “separatista” de pesquisa com a qual também não me alinho. São notáveis a sua atitude de combatente e a sua insatisfação diante dessa situação, ao afirmar que:

> ‘‘poucos ramos do conhecimento sofrem mais do que a Lingüística pela separação entre ‘ciências’ e ‘humanidades’, que <u>ainda</u> se

> mantém nos currículos da maioria das nossas escolas e universidades''.[20]

Assim, com sua visão privilegiada, Lyons constava, já naquela ocasião, que a Teoria Lingüística:

> ''aproveita-se, ao mesmo tempo e, *grosso modo*, eqüitativamente, da abordagem mais tradicional da língua – que é a característica das ''humanidades'' – e da abordagem mais ''científica'' que se desenvolveu recentemente em conexão com os progressos que se verificaram na Lógica Formal, na Análise Computacional e na Teoria dos Autômatos.''

Notável é também constatar que existem projetistas de PLN que se preocupam com o embasamento lingüístico dos estudos sobre o PLN. Winograd (*op. cit.*: 41), nos anos 70, já alertava para o perigo:

> "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."

Para concluir estas reflexões, apresento minhas conclusões sobre esse quadro de desencontros.

A tímida interação entre os "engenheiros da linguagem" e os "cientistas da linguagem", alimentada por tantos desencontros e por certo descrédito mútuo, evidencia:

- O desconhecimento que o pesquisador de uma área demonstra ter do trabalho desenvolvido na outra,

---

20 Grifo meu.

- As tentativas frustradas de compreensão de metalinguagens, conceitos, métodos e técnicas específicos,
- A disseminação de imagens esteriotipadas,
- O incentivo a "lutas teóricas", intra e interdisciplinares…

Enquanto lingüistas lutam para introduzir recursos da informática e das Ciências da Computação em suas pesquisas lingüísticas, projetistas de PLN lutam para domar as línguas naturais, sem poder contar com o auxílio de fundamentação lingüística adequada.

Essa duplicação desnecessária de esforços não só dificulta a descoberta de soluções, que seguramente seria agilizada com o incentivo de trabalho solidário, como é também bastante reveladora, pois deixa transparecer, e até mesmo justificar, a autonomia que se instala entre os Estudos do PLN e os Estudos da Linguagem.

A desvinculação das duas áreas é, porém, extremamente preocupante, porque contribui para aumentar ainda mais os desencontros e, principalmente, minimizar a importância do papel dos lingüistas na proposição e no desenvolvimento de projetos de PLN, que, em geral, resultam de iniciativas tomadas por não-lingüistas e tornam-se privilégio de instituições ou departamentos que investem na pesquisa tecnológica.

O desencontro entre pesquisadores é causa de muitas das inadequações detectadas nos estudos sobre o PLN e, sobretudo, é desestímulo para iniciativas voltadas para a criação de tecnologias lingüísticas.

Cooperar é preciso

Diante desse quadro, há que se reconhecer que o ideal de se construir um universo de discurso comum, que possibilite o diálogo entre os "cientistas da linguagem" e os "engenheiros da linguagem" e, a partir dele, encontrar estratégias de trabalho interdisciplinar que viabilizem a criação de grupos de pesquisa com esse perfil, transforma-se em uma tarefa bastante complexa. Mas por outro lado, acredito, como STAROSTA (1991: 195), que:

> "A cooperação harmoniosa entre ciência e engenharia da linguagem deverá, com certeza, produzir tanto teorias quanto aplicações melhores…"

Desse modo, acredito que tanto os estudos do PLN quanto os Estudos Lingüísticos poderão, de fato, se beneficiar mutuamente com o trabalho interdisciplinar, trabalho que incentive o envolvimento dos vários especialistas na busca conjunta de soluções integradas. Pesquisas dessa natureza poderão contribuir para a proposição de modelos lingüísticos mais completos, explícitos e operacionais e, conseqüentemente, mais apropriados para receberem um tratamento computacional.

Por modelos mais completos, entendo modelos de análise e de descrição que, além de desenvolver cada um dos estratos que compõem a gramática, desenvolva também, ou pelo menos preveja, meios de inter-relacionar um modelo da competência com um modelo do desempenho. Por modelos mais explícitos, entendo modelos formulados em termos de linguagens formais apropriadas; por modelos operacionais, quero enfatizar a necessidade de se construir modelos que possam ser transformados em programas de computador específicos para o PLN.

Assim, será também possível conceber sistemas de PLN mais sofisticados, instrumentos que nos venham a auxiliar na descoberta de novas técnicas e estratégias de trabalho que poderão, por sua vez, ser empregadas na elaboração de análises e descrições cada vez mais precisas e globalizantes dos próprios fenômenos da linguagem.

O desafio é, portanto, viabilizar a formação de um campo de pesquisas acadêmicas interdisciplinares com grande potencialidade tecnológica, integrando recursos teóricos e técnicas de investigação desenvolvidos no âmbito de um conjunto de disciplinas matrizes entre as quais a teoria lingüística que, indiscutivelmente, desempenha papel fundamental.

Harlow & Vincent (1989), ao apresentarem o panorama atual da Lingüística Chomskiana, chegam a prever que as teorias lingüísticas que se preocuparem com as aplicações computacionais de seus resultados poderão ter maior reconhecimento por parte das agências financiadoras e, conseqüentemente, poderão ter também um destaque maior em relação às teorias que não se preocuparem com esse aspecto mais “pragmático” das pesquisas lingüísticas.

Rich (1985), por sua vez, mostra que tanto a Inteligência Artificial quanto as Humanidades podem se beneficiar com o estudo do PLN, uma vez que o problema de simular o complexo competência-desempenho lingüístico humano em uma máquina nos obriga a explicitar ao máximo os elementos constitutivos das línguas naturais, seus princípios e regras.

Logo, é essencial reconhecer a importância de contribuições recíprocas, quer no eixo teoria-aplicação, quer no eixo da

interdisciplinaridade, que visa a minimizar a dicotomia teoria/prática e a fomentar as evidentes contribuições recíprocas de áreas distintas.

CAPÍTULO 2 – A natureza lingüístico-tecnológica do PLN

> *"Natural language processing has a short history. What started out with string manipulation now includes ambitious attempts at simulation of complex linguistic behavior. Yet, it is only during the last 5-10 years that computation has become a concern to linguists…"*
>
> Per-Kristian Halvorsen (1989: 216)

A importância dos estudos lingüísticos para o PLN

A amplitude e a heterogeneidade das pesquisas, somadas à variedade de interesses e à diversidade de métodos empregados, tornam a apreciação histórica da evolução dos estudos sobre o PLN uma tarefa difícil. É possível, porém, estabelecer uma ancoragem histórica que nos permite resgatar momentos decisivos que evidenciam a necessidade de enfrentarmos a complexidade do PLN munidos de conhecimentos interdisciplinares e, em particular, dos conhecimentos específicos que os Estudos Lingüísticos vêm acumulando sobre a linguagem humana.

Para isso, tomo como referencial a Tradução Automática, marco inicial do uso do computador para a investigação das línguas naturais e síntese das questões essenciais do PLN. Além disso, segundo a veemente avaliação de Wilks (1990: 564), a tradução automática é "o espectro que retorna para aterrorizar" os trabalhos sobre o PLN, porque obriga os projetistas a enfrentarem os problemas sem subterfúgios e, segundo Nirenburg *et al.* (1992: 2), trata-se de um "retorno de

vingança", porque apesar das inúmeras críticas constitui, hoje, um dos campos de pesquisa mais atuantes no âmbito do PLN.[21]

As primeiras investigações institucionalizadas do PLN começaram no início da década de 50, depois da distribuição de duzentas cópias de uma carta, conhecida como *Weaver Memorandum*, escrita por Warren Weaver, então vice-presidente da Fundação Rockfeller e exímio conhecedor dos trabalhos sobre a criptografia computacional.[22] Nessa carta, divulgada em 1949, Weaver convidava, universidades e empresas, interessados potenciais, a desenvolverem projetos sobre um novo campo de pesquisa que ficou conhecido como ''Tradução Automática'', ''Tradução Mecanizada'' ou simplesmente MT (abreviatura do inglês "***M****achine* ***T****ranslation*").

Tal documento, embora de caráter predominantemente estratégico, já continha as primeiras preocupações teóricas e metodológicas sobre alguns aspectos importantes que deveriam ser considerados ao se enveredar por esse campo de estudos. Weaver assinalava, por exemplo, a necessidade de se estudar a problemática da polissemia das unidades lingüísticas, o substrato lógico da estrutura das línguas e os universais lingüísticos. Essas questões, entretanto, não estavam no centro das discussões. Como traduzir não era diferente de decifrar códigos, dava-se destaque especial à criptografia, técnica que, hoje, sabemos ser absolutamente inadequada ao tratamento computacional das línguas naturuais.

---

21 Os dados factuais fundamentam-se em Pylyshyn *et al.* (*op. cit.*), Hearn *et al.* (1980), Barr & Feigenbaum (1981), Gardner *et al.* (1981), Berwick (1987), Slocum (1985 e 1989), Ballard & Jones (1990), Kurtzweil (1990), Wilks (*op. cit.*) e Nirenburg *et al.* (*op. cit.*).

22 Essa técnica estava sendo empregada com grande sucesso na tarefa de decifrar códigos das mensagens alemãs durante a Segunda Grande Guerra.

Nos dois primeiros anos após a divulgação da carta de Weaver, as pesquisas sobre tradução automática passaram a ser levadas a sério em importantes instituições norte-americanas como, por exemplo, o Instituto de Tecnologia de Massachusetts (MIT), a Universidade da Califórnia, a Universidade de Harvard e a Universidade de Georgetown. Entre os tópicos mais debatidos estavam as análises morfológica e sintática, a questão da necessidade da pré e pós-edição de textos, a resolução do problema da homografia, técnicas de automatização do processo de consulta a dicionários e a proposição de uma "interlíngua", caracterizada em termos de um sistema de representação abstrata do significado.

A primeira reunião científica sobre tradução automática ocorreu no MIT, em 1952, e a primeira demonstração para o grande público, dois anos depois, na Universidade de Georgetown. A demonstração consistiu em apresentar um sistema capaz de traduzir, do russo para o inglês, 50 frases selecionadas de um texto sobre química. O dicionário construído continha 250 palavras e a gramática escrita para o russo possuía apenas seis regras. O sucesso desse experimento acabou atraindo a atenção de várias instituições financiadoras nos Estados Unidos e em outros países, principalmente na então União Soviética.

Houve várias tentativas de se estender essa experiência bem-sucedida para cobrir um maior número de estruturas e de itens lexicais de um número maior de línguas que ser revelaram, entretanto, elas estavam muito aquém do esperado pelas agências financiadoras.

É importante ressaltar, neste ponto da história do PLN, que as considerações e motivações que levaram as agências financiadoras a

estimular os estudos sobre a tradução automática, e que deram o tom das pesquisas até o início dos anos 60, eram basicamente as seguintes:

[i] numa época em que começava a haver uma explosão de informações, a automação do processo de tradução, em princípio, significaria mais eficiência e era, antes de tudo, vista como um negócio importante e lucrativo;

[ii] acreditava-se também que projetar e implementar um modelo de tradução automática eram tarefas relativamente simples, bastando criar dicionários informatizados e programas que se incumbissem, apropriadamente, de executar "as consultas";

[iii] esperava-se que as necessárias consultas aos dicionários fossem drasticamente reduzidas, à medida que tais dicionários fossem sendo implementados;

[iv] considerava-se a criptografia uma técnica eficiente e apropriada para a execução da tarefa.

Para se ter uma idéia da má qualidade da tradução gerada pelos primeiros sistemas de tradução automática, impulsionados pelas motivações descritas acima, basta tentar ler o segmento "traduzido" do russo para o inglês, extraído de Barr & Feigenbaum (*op. cit.*: 235):

> (In, At, Into, To, For, On) (last, latter, new, latest, lowest, worst) (time, tense) for analysis and sinthesis relay-contact electrical (circuit, diagram, scheme) parallel-(series, successive, consecutive) consistent (connection, junction, combination) (with,from) (success, luck) (to be utilize, to be take advantage of) apparatus Boolean algebra.

Como se observa, esses sistemas simplesmente listavam as várias possibilidades de tradução de cada palavra encontrada no texto de origem. Nenhuma tentativa de análise sintática era cogitada. A grande

maioria das ‘‘traduções’’ feitas pela máquina eram, conseqüentemente, de péssima qualidade e exigiam constantes revisões por parte de tradutores humanos.

Bar-Hillel foi o maior crítico dos trabalhos produzidos nessa pré-história da Tradução Automática. Sua principal crítica dizia respeito à própria possibilidade de se conseguir criar sistemas com essa sofisticação. Para ele, uma tradução exclusivamente automática e de qualidade era absolutamente impossível.

Devido à sua reputação de grande conhecedor das pesquisas sobre o tema, Bar-Hillel, com suas severas críticas, além de silenciar muitas iniciativas, incentivou a divulgação, em 1964, de um relatório fulminante, contendo uma avaliação negativa do nível das pesquisas até então produzidas. Esse relatório, elaborado pelo Comitê Assessor de Processamento Automático das Línguas Naturais (*Automatic Language Processing Advisory Committee* - ALPAC),[23] concluía que, até aquele momento, não só não havia tradução automática de texto científico algum, como também não havia perspectiva alguma de viabilidade desse tipo de empreendimento, principalmente porque a necessidade constante de contratação de pessoal especializado em tradução para realizar as tarefas de pré e pós-edição dos textos tornava a tradução automática um empreendimento absolutamente inócuo. Como conseqüência, as agências financiadoras norte-americanas e britânicas reduziram drasticamente seus incentivos.[24] O reflexo imediato dessa decisão foi o

---

23 Esse comitê foi criado em 1964 pela Academia Nacional de Ciências dos Estados Unidos para elaborar um relatório avaliativo para ser apresentado aos principais órgãos intitucionais norte-americanos envolvidos em projetos de tradução automática: o Departamento de Defesa, a Agência Central de Inteligência dos Estados Unidos e a própria Academia Nacional de Ciências.

24 Essa atitude não chegou, entretanto, a atingir as pesquisas que estavam sendo desenvolvidas no Canadá, na então União Soviética, na França, na Alemanha e na Itália.

desaquecimento das pesquisas nesse campo e, conseqüentemente, dos projetos que visavam à criação de sistemas com finalidades comerciais.

Além desse documento, trabalhos de pouco interesse lingüístico também contribuíram para o descrédito de pesquisas sobre a tradução automática e, de maneira geral, sobre todo o campo do PLN. Contar, por exemplo, quantas vezes a palavra ‘‘*king*’’ ocorria em obras de Shakespeare era considerado um estudo sobre o PLN. O propósito desse tipo rudimentar de análise era verificar se a autoria de um determinado texto podia realmente ser atribuída a um determinado autor. Primeiro, calculavam-se estatísticas de certas palavras freqüentemente encontradas nos textos que eram indiscutivelmente do autor analisado. Depois tomavam-se as estatísticas das mesmas palavras nos textos em que se queria comprovar (ou não) a autoria. A comparação dos resultados podia dar pistas sobre a questão da autoria.

Depois de muitas experiências negativas e concepções equivocadas em relação ao tratamento computacional das línguas naturais, a partir de meados da década de 70, os trabalhos de tradução automática foram retomados com uma atitude mais acadêmica e realista. Além disso, há que se reconhecer que o relatório da ALPAC acabou por penalizar muitos projetos sérios de tradução automática que caminhavam para o sucesso. Um deles, por exemplo, o sistema piloto GAT (datado de 1962), originado a partir do experimento na Universidade de Georgetown, era capaz de produzir traduções do russo para o inglês de qualidade considerável (NIRENBURG *et al.*, *op. cit.*: 4):

> Automation of the process of a translation, the application of machines, with a help which possible to effect a translation without a knowledge of a corresponding foreign tongue, would be an important step forward in the decision of this problem.

Desta vez, desvencilhados de interesses bélicos, estratégicos e imediatistas, os pesquisadores passaram a ser mais cautelosos diante do complexo processo de tradução e da própria sofisticação do código lingüístico. Entre os projetos que refletem essa maturidade das pesquisas encontram-se, por exemplo, os sistemas TAUM-METEO, SYSTRAN, ATLAS II , EUROTRA e KBMT, desenvolvidos nas décadas de 70 e 80.

Assim, por causa de experiências bem-sucedidas como o projeto GAT, e, de certa forma, resistindo aos impactos negativos do relatório governamental, outros projetos de PLN, teóricos e aplicados, e não exclusivamente sobre a tradução automática, uma de suas aplicações potenciais, passaram também a ser desenvolvidos. O ímpeto de muitos pesquisadores, que encontravam no PLN um estímulo para o desenvolvimento de pesquisas teóricas, não foi totalmente abalado.

Em 1970, um desses militantes, Terry Winograd, em sua tese de doutorado no MIT, criou um sistema computacional que passou a ser o marco dos estudos acadêmicos sobre o PLN: o sistema SHRDLU, também conhecido como “Mundo dos blocos”. Com esse trabalho, Winograd conseguiu mostrar para a comunidade científica que a interação homem-máquina por meio de línguas naturais era realmente possível.

O sistema proposto por Winograd simula, sob forma de representação gráfica no monitor do computador, o braço de um robô que manipula um conjunto de blocos sobre a superfície de uma mesa, executando as “ordens”, que lhe são digitadas, em inglês, no teclado. No monitor, vê-se o braço do robô executando o que lhe é solicitado. A

seguir, apresento um segmento de uma sessão com o robô do mundo dos blocos (WINOGRAD, *op. cit.*: 8-15), adaptado para o português:[25]

SITUAÇÃO INICIAL: **Há um bloco vermelho sob um bloco verde**.
O usuário digita:
Pegue o bloco vermelho grande.
O robô retira o bloco verde, encontra um local para colocá-lo, apanha o bloco vermelho e escreve na tela do computador:
OK.

SITUAÇÃO 2: **Há uma pirâmide azul dentro de uma caixa**.
O usuário digita:
A pirâmide verde é minha.
O robô não se move e escreve na tela do computador:
OK.

SITUAÇÃO 3: Há três pirâmides sobre a mesa.
O usuário digita:
Pegue a pirâmide.
O robô não se move e escreve na tela do computador:
NÃO COMPREENDO.QUAL PIRÂMIDE VOCÊ QUER QUE EU PEGUE?

SITUAÇÃO 4: Há um bloco que é o maior de todos.
O usuário digita:
Denomine o maior boco de "superbloco".
O robô não se move e escreve na tela do computador:
OK.

SITUAÇÃO 5: O bloco maior é o bloco vermelho referido na SITUAÇÃO 1.
O usuário digita:
Você já pegou um superbloco desde que começamos a sessão?
O robô não se move e escreve na tela do computador:
SIM.

SITUAÇÃO FINAL: O robô nunca derruba blocos e o usuário o acusa de derrubar um bloco.
O usuário digita:
Por que você o derrubou?
O robô não se move e escreve na tela do computador:
SUPONHO QUE, POR "O", VOCÊ ESTEJA SE REFERINDO AO SUPERBLOCO. NÃO POSSO EXPLICAR UM EVENTO IMPOSSÍVEL.

Com esse programa, ficou demonstrado para a comunidade acadêmica que a máquina, mesmo de modo primitivo, poderia ser programada para processar objetos linguísticos, ou seja: interpretar perguntas, asserções e comandos; inferir, explicar ações e, até mesmo, "aprender" palavras novas. A partir de Winograd, o PLN passou a constituir, de fato, um objeto relevante para a pesquisa científica.

25 No sexto capítulo será feita uma descrição esquemática desse programa.

Conseqüentemente, uma multiplicidade de investigações científico-acadêmicas passaram a se somar às pesquisas tecnológicas que dominavam o campo.

Para finalizar, apresento uma síntese da evolução dos estudos sobre o PLN em termos do grau de sofisticação lingüística alcançado:

DÉCADA DE 50 : A TRADUÇÃO AUTOMÁTICA

- Sistematização computacional das classes de palavras descritas nos manuais de gramática tradicional;
- Identificação computacional de poucos tipos de constituintes oracionais.

DÉCADA DE 60: AS 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*.

DÉCADA DE 70: A 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.

DÉCADA DE 80: A SOFISTICAÇÃO DOS SISTEMAS

- Desenvolvimento de teorias linguísticas motivadas pelos estudos do PLN como, por exemplo, a *gramática sintagmática generalizada* e a *gramática léxico-funcional*.

DÉCADA DE 90: OS SISTEMAS BASEADOS EM "REPRESENTAÇÕES DO CONHECIMENTO"

- Desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos linguísticos e extralinguísticos e das estratégias de inferência envolvidos nos processos de

produção, manipulação e interpretação de objetos linguísticos para os quais os sistemas são projetados.

RUMO AO SÉCULO XXI: OS SISTEMAS "HÍBRIDOS

– Desdobramentos e consolidação dos conhecimentos da década de 90, acrescidos de iniciativas devdem visar à integração de metodologias calcadas em métodos estatísticos avançados, na manipulação de imensos corpora de textos e na aplicação de redes neurais.

Após 25 anos de refinamentos, o estudo do PLN, enquanto área multidisciplinar, hoje, é uma realidade. Um dos centros de pesquisa norte-amareicanos mais representativos, refletindo a filosofia de se desenvolver trabalhos cooperativos entre Lingüística, Inteligência Artificial e Ciência da Computação é o *Center for the Study of Language and Information* (CSLI).[26] Na Universidade Carnegie Mellon, outro centro norte-americano de destaque,[27] vários projetos de vanguarda encontram-se atualmente em desenvolvimento: SPHINX,[28] projeto de síntese e reconhecimento da fala; DOC-PAT COMMUNICATION, projeto de elaboração de interfaces em lingua natural que auxiliam a triagem de pacientes através de entrevistas geradas por computador; DIOGENES (NIRENBURG *et al.*, *op. cit.*), projeto de geração de texto escrito; CLARIT (EVANS *et. al.*, 1991), projeto de recuperação de informação contida em grandes bases de textos; KBMT (NIRENBURG *et al.*, *op. cit.*), projeto de tradução automática, entre outros.[29]

---

26 Fundado em 1983, esse centro avançado de pesquisa é uma iniciativa conjunta da Universidade de Stanford, do Centro Internacional de Stanford e do Centro de Pesquisa da Xerox em Palo Alto, California, EUA (HALVORSEN, *op. cit.*: 198).

27 Segundo relatório elaborado pela NASA (*cf.* GEVARTER, *op. cit.*: 123)

28 *Cf. Areas of CL Research at CMU*, (EVANS, 1990).

29 Gevarter (1984) fornece uma listagem das instituições norte-americanas envolvidas em projetos de PLN.

Destacam-se estes periódicos de divulgação dos trabalhos da área: *Computational Linguistics* (o antigo *American Journal of Computational Linguistics*), *Artificial Intelligence*, *Canadian Journal of Artificial Intelligence* e *Cognitive Science*. Além disso, há os encontros científicos regulares: as reuniões anuais da *Association for Computational Linguistics* (ACL), as conferências patrocinadas pela *Americam Association for Artificial Intelligence* (AAAI) e as conferências internacionais bienais *International Conference on Computational Linguistics* (COLING) e a *International Joint Conference on AI* (IJCAI) (BALLARD & JONES, *op. cit.*: 146).

Um laboratório em ebulição

O levantamento de trabalhos revela que o PLN é um "laboratório em ebulição". Sistemas computacionais são projetados, estudados, implementados, testados e comercializados, uma vez que a indústria de informática cresce assustadoramente. Com graus diferentes de sofisticação lingüística, as possibilidades de aplicação do estudo do PLN na construção de **Sistemas de PLN** (SPLN) são expressivas e impressivas. Destacam-se:

- Sistemas de manipulação de bases de dados;
- Sistemas tutores, ou seja, sistemas de estudo assistido por computador;
- Sistemas de automação de tarefas administrativas e gerenciais;
- Sistemas programação automatica de computadores;
- Sistemas de processamento automático de textos e informações;

- Sistemas especialistas;
- Sistemas de tradução automática;
- Sistemas científico-acadêmicos.

Sistemas de manipulação de bases de dados – Nos sistemas de manipulação de base de dados, o papel do SPLN é servir de módulo de comunicação entre o usuário e a base de dados, "traduzindo" frases-instruções, isto é, instruções codificadas em frases, digitadas em um terminal, para a linguagem específica do sistema de gerenciamento de dados que, por sua vez, se encarrega de manipular as informações.[30] Esses SPLNs são genericamente denominados "sistemas de perguntas e respostas" (WEBBER, 1990). Exemplos significativos são os seguintes sistemas: BASEBALL (GREEN *et al.*, 1986) – que responde a perguntas sobre o mês, o dia, o local, os times e os resultados referentes aos jogos da Liga Americana de Baseball; RENDEZVOUS (BALLARD & JONES, *op. cit.*: 138) – que auxilia o usuário a encontrar informações em uma base de dados que registra o estoque de uma empresa, reconhecendo qualquer tipo de frase, fragmentada ou não, gramatical ou não, e apenas descarta frases que reportam a entidades fora do domínio do discurso estabelecido; LIFER (BARR & FEIGENBAUM, *op. cit.*: 316-21) – que auxilia implementadores de sistemas na criação do próprio SPLN; PLANES&JETS (BALLARD & JONES, *op. cit.*: *loc.cit.*) – que, além de se comunicar com o usuário por meio de frases, possui um dispositivo adicional que monitora a comunicação entre o usuário e o sistema, permitindo-lhe otimizá-la; LUNAR (WOODS, 1978) – que é capaz de

30 O sistema de gerenciamento de dados é um programa específico, escrito em uma linguagem de programação convencional, que se encarrega de efetuar a organização, catalogação, localização, armazenamento, recuperação e manutenção das informações de uma base de dados.

interpretar vários tipos de frases durante o processo de consulta a informações sobre a geologia de rochas lunares; e TEXT (McKEOWN, 1985) – que gera textos da extensão de parágrafos como respostas à solicitação de informação sobre os veículos aquáticos da marinha americana.[31]

Sistemas tutores – Há basicamente dois tipos de sistemas de estudo por computador. Os sistemas considerados tradicionais (*Computer-Aided Instruction*)[32] e os sistemas chamados inteligentes (*Intelligent Computer-Aided Instruction*).[33] Nos sistemas tradicionais, os conteúdos são estruturados de maneira fixa e apresentados no monitor em forma de instrução programada e ramificada, previamente especificadas pelo projetista do sistema. O módulo lingüístico fica reduzido à manipulação de estruturas lingüísticas pré-formatadas. Por esse motivo, esses sistemas são de pouco interesse do ponto de vista do PLN.

Nos sistemas inteligentes, por outro lado, o SPLN desempenha papel essencial. Os conteúdos são estruturados em termos de "redes de conhecimentos", compostas de fatos, regras e relações que permitem ao sistema desencadear um "diálogo socrático" com o aluno, simulando a situação em que aluno e professor discutem tópicos específicos de conteúdo. É preciso munir a máquina de um

---

31 É importante esclarecer que uma simples mensagem de erro, emitida por um programa como resposta a algum tipo de falha do sistema computacional, não é considerada uma produção de texto. Uma mensagem de erro não significa nada para o sistema. Trata-se de um texto pré-escrito pelo programador. Mesmo que as mensagens fossem parametrizáveis, isto é, possuíssem variáveis para serem preenchidas por nomes de indivíduos ou objetos diferentes, por exemplo, tais mensagens também não seriam consideradas textos gerados pelo computador.

32 *Cf.* Farghaly (1989).

33 *Cf.* Bailin *et al.* (1988), Bailin (1989) e Bailin & Levin (1989).

conhecimento lingüístico altamente sofisticado para que ela possa simular o “diálogo socrático”.

Os sistemas tutores inteligentes destacam-se pela riqueza de pesquisas que geram, já que permitem ao pesquisador desenvolver simulações diversas: modos de ensinar os conteúdos, de representar o processo de aprendizagem, de caracterizar o aluno-usuário, de analisar, corrigir e comentar erros, de avaliar o aprendizado, de fazer com que o sistema antecipe dúvidas, modifique suas estratégias de ensino e melhore sua interação com o aluno, entre outras. Alguns exemplos ilustram algumas iniciativas: SCHOLAR (BALLARD & JONES, *op. cit.*: 139, 141) – que não se limita a oferecer respostas já armazenadas no sistema, mas ‘‘analisa’’ a situação do diálogo e escolhe a melhor resposta para aquele momento da interação; STUDENT (BOBROW, 1968) – que auxilia o aluno na resolução de problemas de álgebra elementar formulados em inglês. ALICE (EVANS & LEVIN, 1990) – protótipo de sistema tutor de estudos de língua estrangeira no qual se destacam as seguintes características: seu SPLN é capaz de executar análises morfológicas e sintáticas, gerar frases simples em inglês, francês, espanhol, alemão e japonês e contextualizá-la por meio de textos e imagens.

Sistemas de automação de tarefas administrativas – Esses sistemas auxiliam nas tarefas de rotina de setores administrativos e gerenciais. SCHED é um programa capaz de gerenciar agendas de reuniões; GUS (BOBROW, 1986) fornece informações sobre planejamento de viagens aéreas; UC responde perguntas sobre o ambiente computacional UNIX; VIPS seleciona e manipula objetos no monitor do computador por meio de comandos orais; CRITIQUE detecta

erros ortográficos e gramaticais e analisa palavras, sintagmas e frases que possam comprometer a leitura fluente de documentos administrativos.[34]

Sistemas programação automatica de computadores [35] – Esses sistemas são projetados com a finalidade de facilitar a interação entre o programador e a máquina. A estrutura desses sistemas é bastante complexa, pois deles são exigidas inúmeras tarefas: receber e organizar a informação dada pelo programador, fornecer os elementos de programação, coordenar os procedimentos de síntese dos programas a serem gerados e, finalmente, gerar um programa aceitável. Para executar essas tarefas, o sistema desencadeia uma "entrevista" com o programador, na qual o sistema adquire um modelo dos processos computacionais necessários, verifica a sua correção, seleciona as estruturas de dados apropriadas para a execução da tarefa solicitada e, por fim, fornece o programa. NLPQ e SAFE são exemplos ilustrativos dessa modalidade.[36]

Sistemas de processamento automático de textos e informações – Depois de agrupar relatórios de exames radiológicos e

---

34 Os sistemas SCHED, UC, VIPS e CRITIQUE encontram-se descritos em Ballard & Jones (*op. cit.*: 140).

35 Biermann (1990) concebe a programação de computadores como um processo de construção de um código executável pela máquina a partir de informações fragmentadas. Essas informações são, em geral, de natureza diversa: idéias vagas sobre o que se espera do programa, sobre o tipo de dados que deverá alimentar o sistema, o tipo de algoritmo a ser utilizado ou exemplos do funcionamento pretendido. O produto final da programação consiste em uma seqüência de códigos capaz de receber informações específicas de um determinado domínio e processá-las, produzindo, como resultado, outras informações. A programação é uma atividade de programadores. Trabalhos têm sido propostos, porém, para que, pelo menos, parte dessa atividade seja executável pela própria máquina. Essa nova perspectiva abre, então, a possibilidade de metodologias, dentre as quais está o desenvolvimento de sistemas de programação automática a partir da interação em linguagem natural entre homem-máquina. Essa metodologia envolve a "tradução" de descrições em linguagem natural em especificações formais e, portanto, programáveis.

36 *Cf.* Bierman (*op.cit.*: 32)

convertê-los no formato de uma base de dados, esse tipo de sistema possibilita ao usuário consultar informações por meio de perguntas. A informação de entrada e saída do sistema é codificada em frases que são analisadas e sintetizadas, segundo um padrão pré-estabelecido. Esse padrão, definido a partir de características sintáticas das palavras, é armazenado sob a forma de uma tabela em que cada coluna contém uma parcela da informação necessária para a interpretação da frase-pergunta e para a construção da frase-resposta. (*cf.* GRISHMAN, *op. cit.*: 151-53).

Sistemas especialistas [37] – O livro é, sem dúvida, o meio de registro e armazenamento de conhecimentos mais difundido de que dispomos. Os conhecimentos nele armazenados, entretanto, têm um caráter passivo. Sua aplicação na resolução de problemas depende necessariamente de um agente humano capacitado para recuperá-los, interpretá-los e decidir como explorá-los de maneira apropriada.

Os programas de computadores convencionais, apesar de serem capazes de manipular informações segundo esquemas lógicos de decisão, não são suficientemente sofisticados para simular um agente humano naquelas tarefas. Um programa convencional é basicamente constituído de duas partes distintas: algoritmos e dados. Os algoritmos determinam como resolver os problemas, e os dados caracterizam os parâmetros envolvidos no processo.

Como grande parcela das informações "geradas e processadas" pelo homem é constituída de uma pluralidade de informações fragmentadas, é preciso criar novos esquemas de decisão, capazes de organizar os fragmentos em um todo coerente.

---

37 *Cf.* Hayes-Roth (1990).

Para preencher essa lacuna, criam-se os sistemas especialistas, que são projetados para utilizar parcelas do conhecimento humano no processo de **resolução de problemas**.[38] Nesses sistemas, são implementados mecanismos de aquisição, representação e implementação desse conhecimento, o que os tornam mais eficientes que os meios mais convencionais de armazenamento, manipulação e transmissão de informações. Projetados com esquemas complexos de decisão, os sistemas especialistas são capazes de agrupar fragmentos de informação numa base de dados e sobre ela operar segundo regras de inferência bastante complexas. A estrutura, o modo de incorporação da informação e o impacto que seu funcionamento causa sobre o usuário, que tem a ilusão de estar interagindo com um interlocutor inteligente, são características que os tornam diferentes dos sistemas convencionais.

Encontramos sua aplicação na resolução de problemas em áreas como diagnóstico médico, conserto de equipamentos, configuração de computadores, interpretação de dados e estruturas químicas, interpretação de imagens e da linguagem oral, interpretação de sinais, sistemas de planejamento e consultoria, entre outras. Destacam-se:[39] DENDRAL, o primeiro sistema especializado, criado para ajudar os químicos a determinar a estrutura molecular; MYCIN – que incorpora 400 regras heurísticas escritas em inglês para diagnosticar doenças sangüíneas infecciosas, oferecendo explicações sobre as conclusões ou perguntas por ele geradas; INTERNIST – que contém 100.000

---

38 A **resolução de problemas** é o principal fenômeno estudado pela Inteligência Artificial. Seu estudo consiste na investigação dos sistemas computacionais que são projetados para desencadear processos que envolvem a descoberta, ou a construção, de soluções de problemas. Entre os objetos analisados, encontram-se: a caracterização, a classificação, a formulação e a representação do problema e os procedimentos necessários para a sua resolução. (*Cf.* RICH, 1983 e 1985; WINSTON, 1984).

39 *Cf.* Hayes-Roth (*op. cit.*).

julgamentos sobre relações entre doenças e sintomas; HEARSAY-II – que combina sistemas especializados múltiplos na tarefa de interpretar segmentos conexos de fala a partir de um léxico contendo 1.000 palavras; e XCOM – que incorpora 1.000 regras de implicação lógica para executar a tarefa de configuração dos componentes de um computador VAX.

Sistemas de tradução automática [40] – Os sistemas de tradução automática podem ser classificados de acordo com a metodologia de tradução empregada: sistemas diretos, sistemas transferenciais e sistemas interlinguais.

Os sistemas diretos buscam correspondências diretas entre as unidades lexicais da língua de partida e da língua de chegada como, por exemplo, o sistema SYSTRAN, criado para traduzir relatórios sobre a missão espacial Apollo-Soyuz.

Os sistemas de transferência já são mais sofisticados como, por exemplo, o sistema TAUM-METEO, que até hoje traduz os relatórios meteorológicos do Canadá do inglês para o francês; e EUROTRA, que pretende traduzir as línguas dos países pertencentes ao Mercado Comum Europeu. Estes sistemas efetuam a análise sintática da frase da língua de partida e, através de regras de transferência sintática, constroem a representação sintática da frase da língua de chegada.

Os sistemas interlinguais são os mais sofisticados dos três como, por exemplo, os sistemas ATLAS-II, PIVOT, ULTRA e KBMT-89, nos quais a língua de partida e a língua de chegada são intermediadas por uma *interlíngua*, isto é, uma representação abstrata universal do

40 *Cf.* Slocum (1985 e 1989), Marín (1989), Wilks (1990) e Nirenburg *et al.* (*op. cit.).*

significado para a qual a língua de partida é ‘‘traduzida’’ e, a partir da qual, a língua de chegada é ‘‘gerada’’.

Sistemas científico-acadêmicos - Schank & Riesbeck (*op. cit.*: 3-8), desde 1975, vêm projetando uma série de programas para testar sua teoria chamada *dependência conceitual*, que contém os conceitos de *frames*, *scripts*, *planos* e *metas*. Criaram o programa MARGIE para testar a sua teoria e mostrar a viabilidade de se criar uma linguagem de representação semântica em termos de uma interlíngua, independente de qualquer língua em particular. Composto de um analisador conceitual, que transforma as frases de entrada em uma representação conceitual, um gerador de frases e um mecanismo de inferências (tradução do inglês *inference engine*), esse programa executa dois tipos de operações sobre frases: paráfrase e inferência. No modo paráfrase, dada uma frase como *John killed Mary by choking her*, o programa gera paráfrases como *John strangled her* e *John choked Mary and she died because she was unable to breathe*. No modo inferência, dada uma frase como *John gave Mary an aspirin*, o programa gera as seguintes inferências: *John believes that Mary wants an aspirin*, *Mary is sick*, *Mary wants to feel better* e *Mary will ingest the aspirin*. Os sistemas SAM e PAM, uma evolução de MARGIE, foram desenvolvidos para simular a compreensão automática de pequenas histórias.

Merecem destaque outros sistemas acadêmicos. Raphael (1968) desenvolveu o programa SIR, implementando mecanismos de inferência mais sofisticados que os sistemas anteriores. Esse programa simula relações do tipo conjunto-subconjunto, parte-todo e possuidor-possuído. Marcus (1980) desenvolveu um analisador gramatical que constrói a estrutura sintática de frases com base na teoria chomskiana.

Cullingford (1981), tomando por base os conceitos de *scripts* e *dependência conceptual*, propôs um sistema de processamento de textos que ‘‘lê’’ artigos de jornais, a partir dos quais, produz *scripts* representando tanto o seu conteúdo explícito quanto o implícito. Berwick (1985), em sua tese de doutorado, fundamentando-se também na teoria gramatical chomskiana, propôs um modelo que simula a aquisição de componentes de uma gramática: categorias lexicais, regras sintáticas e as posições relativas de determinantes e modificadores em relação ao núcleo do sintagma. Bresnan (1987) e sua equipe projetaram uma plataforma computacional para o desenvolvimento de gramáticas para diferentes línguas com base na Teoria da Gramática Léxico-Funcional, a ser esquematicamente apresentada no quarto capítulo.

A essência lingüística e tecnológica do PLN

Nesse emaranhado de pesquisas, encontra-se dispersa a concepção lapidar de Winograd (*op. cit.*: 1) para o PLN. Nela, encontram-se os elementos essenciais para o desenvolvimento do empreendimento e, sobretudo, a indispensável ancoragem lingüística:

> “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 deve possuir um modelo simples de sua própria mentalidade. Ele deve 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 representadas dentro do sistema.”[41]

Além de evidenciar o complexo de conhecimentos e habilidades envolvido no processo de comunicação verbal, e que precisam estar representados dentro de um SPLN, Winograd (*op. cit.*: ix) nos mostra que 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 se quer 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.”[42]

---

41 Grifo meu.

42 Grifo meu.

Perspectivas

Do ponto de vista da pesquisa aplicada, o estudo do PLN visa, em última instância, à implementação de sistemas computacionais em que a comunicação entre o homem e o computador possa ser estabelecida por meio de parcelas de uma língua natural, ou "pseudo-língua natural", e não por meio de instruções e comandos convencionais, codificados em uma linguagem artificial de programação qualquer. Nesse sentido, a pesquisa reveste-se de um caráter tecnológico e transforma-se em um objeto cobiçado pela voraz indústria da informática que, cada vez mais, precisa tornar seu produtos menos "enigmáticos" e mais adaptados às necessidades dos seus clientes.

Assim, projetar SPLNs significa, antes de tudo, tornar os computadores máquinas mais acessíveis, principalmente ao usuário comum, que, ainda hoje, deles se afasta por considerá-los complexos demais ou absolutamente "idiotas" e dispensáveis. Aqueles que vencem esse primeiro impulso são obrigados a se moldar às exigências "das esfinges": memorizar conjuntos de teclas e comandos e aprender a "linguagem das janelas, dos menus e dos ícones".

O usuário mais ousado e o especialista, por sua vez, ainda enfrentam sérios problemas. Não são poucas as vezes que os cientistas da computação, programadores e técnicos em informática são obrigados a ‘‘digerir’’ volumes e mais volumes de manuais técnicos, muitas vezes mal escritos, e a se prostrar diante das mais variadas e complicadas linguagens de programação, além de ter de dominar os diversos sistemas operacionais DOS, WINDOWS, OS/2 e UNIX, projetados por empresas diversas, e ávidas por derrotar umas às outras na conquista pelo mercado.

Criar programas que facilitem a comunicação entre o computador e o usuário, já iniciado no universo da informática, ou não, significa, portanto, desenvolver sistemas computacionais que incorporem algum tipo de SPLN, isto é, um conjunto de programas específicos, integrado ao sistema e projetado para executar a complexa tarefa de interpretar e gerar informações veiculadas por mensagens lingüisticamente construídas. Em outras palavras, estudar o PLN é fornecer subsídios para a implementação de programas computacionais específicos que, de alguma forma, envolvem a manipulação de objetos lingüísticos. Entre os programas computacionais que apresentam essa característica encontram-se, por exemplo, os "corretores ortográficos", os "corretores gramaticais", os "dicionários de sinônimos e antônimos", os "programas de hifenização" e os interessantes programas que convertem grafemas em fonemas, transformando, assim, o computador em uma máquina capaz de "ler" qualquer texto em "voz alta".[43] Essas "aplicações", como preferem chamar os projetistas, são, em geral, construídos para serem integrados a outras como, por exemplo, os diversos processadores de texto.[44]

Ao lado desses programas, cuja sofisticação ainda está consideravelmente aquém da esperada pelos usuários, e infinitamente aquém da grande meta – fazer com que a máquina de fato processe as línguas naturais –, há projetos bastante arrojados sendo desenvolvidos em centros de pesquisa acadêmicos, ligados a universidades, ou em

---

43 O programa *Monologue*™ da Creative Labs, que acompanha alguns kits de multimídia, é um bom exemplo de "programa leitor".

44 Ao lado dos conhecidos aplicativos acoplados ao processador de textos *WORD for Windows*™ da Microsoft, encontram-se também programas mais específicos que fornecem também meios para a correção de pontuação, de emprego inadequado de termos e de estilo e, até mesmo, de alguns tipos de erros gramaticais. O *Gram.mat.ik 5* ™ da Reference Software International, por exemplo, enquadra-se nessa categoria.

centros de pesquisa montados e financiados pela própria indústria de informática.[45] Entre eles, incluo (i) os projetos que vêm estudando as possibilidades de implementação de programas de PLN que servem de interface entre o usuário e uma base de dados,[46] permitindo-lhe manipulá-la por meio de instruções em uma língua natural; (ii) os projetos de desenvolvimento dos sistemas especialistas; e (iii) os projetos voltados para os cobiçados sistemas de tradução automática, ainda distantes de ser uma realidade.

Além de cumprir objetivos mais tecnológicos, estudar o PLN pode significar também desenvolver projetos de caráter acadêmico como, por exemplo, criar modelos computacionais que simulam os processos de produção e recepção de enunciados e textos, ou que sirvam de instrumento no processo de construção e teste de modelos lingüísticos. Dessa perspectiva, o SPLN passa a ser uma plataforma de trabalho para o desenvolvimento de modelos de análise e descrição lingüísticas, na qual os lingüistas e outros pesquisadores, envolvidos com os estudos da linguagem humana *per se*, podem se dedicar à formalização, operacionalização, teste, refinamento e reformulações de seus próprios modelos. Na verdade, a investigação do PLN sugerida nesta tese pode trazer uma nova perspectiva de análise dos fenômenos da própria linguagem humana, uma abordagem que possibilita ao lingüista dissecar os fenômenos da linguagem humana com uma das ferramentas mais sofisticadas que o homem foi capaz de construir – o computador.

---

45 Mesmo os projetos desenvolvidos nas universidades contam com substancial apoio da indústria de informática.

46 Uma base de dados é o conjunto de todos os dados (numéricos, alfanuméricos, gráficos ou sonoros), armazenados no sistema em formato operável pela máquina.

Com isso, não pretendo traçar uma linha divisória rígida entre o que se poderia chamar de “sistemas acadêmicos” e de “sistemas aplicativos”, mas, antes de tudo, chamar a atenção para o leque de possibilidades acadêmicas que pesquisas sobre o PLN abrem para o avanço dos próprios estudos da linguagem e, conseqüentemente, para o desenvolvimento de aplicações mais robustas e lingüisticamente fundamentadas, revelando assim os contornos da face tecnológica dos estudos da linguagem.

Assim, essa multiplicidade de projetos possibilita pesquisas científicas inter e multidisciplinares, com potencialidades tecnológicas, gerando novas perspectivas de trabalho, em particular, para os estudiosos da linguagem que poderão não só participar de projetos como consultores, mas principalmente como proponentes de SPLNs que poderão ser integrados aos mais diversos tipos de sistemas aplicativos.

Há que se ressaltar, neste ponto, que, além da concepção lapidar de Winograd, fatos marcantes, uns relacionados com a capacidade computacional das máquinas, outros resultantes da mudança de concepção de programação e outros ainda decorrentes de sua democratização, contribuíram decisivamente para a proposição de uma estratégia do PLN.

Ao recordarmos que, na década de 40, os computadores eram projetados para processar dados exclusivamente numéricos e que o poder computacional de que dispunham poderia ser comparado ao de uma simples calculadora de bolso, parte de nossos utensílios básicos, há que se reconhecer que foi preciso muita ousadia tentar estudar a tradução automática, que vem se revelando como um dos problemas mais difíceis de ser equacionado. É preciso recordar que os primeiros

computadores eram apenas poderosas máquinas de cálculos numéricos, fato que, de imediato, os colocou a serviço das matemáticas e das engenharias e os transformou em um instrumento, até muito recentemente, dispensável para as ciências humanas.

É preciso também reconhecer que a democratização do uso dos computadores é muito recente. Embora os computadores sejam, sem dúvida alguma, máquinas bastante difundidas e intensamente utilizadas por pesquisadores das ciências exatas desde meados da década de 60, seu acesso era bastante restrito. Por questões físicas, uma vez que a instalação dos primeiros computadores exigia ambientes amplos e apropriados, e técnicos, porque seu manuseio era extremamente complexo, fez-se necessária a criação dos sofisticados Centros de Processamentos de Dados (CPDs). Eram, pois, esses centros que se encarregavam de prestar os serviços computacionais solicitados por matemáticos, físicos e engenheiros que, embora pesquisadores "privilegiados", geralmente nem chegavam perto dos gigantescos *mainframes*, situação que se manteve até o início dos anos 80.

O alto custo, a complexidade de instalação dos *mainframes* e a necessidade de grandes investimentos com pessoal técnico podem ser apontados como fatores que acabaram levando o mercado consumidor de produtos computacionais a optar pela sua total substituição, ou pela aquisição de estações de trabalho e computadores do tipo pessoal. Graças à pressão do mercado consumidor e à descoberta de novas tecnologias, tornou-se possível, hoje, contar com equipamentos mais baratos, computacionalmente mais poderosos, de manipulação mais acessível e, até mesmo, prontos para serem conectados uns aos outros por meio de redes. Com o avanço da tecnologia computacional e com

sua maior democratização, é concebível pensar na criação de modelos de PLN em estações de trabalho e, até mesmo, em modestos computadores pessoais.

Essas facilidades computacionais e, principalmente, a crescente introdução dos computadores no "universo das humanidades" são fatores favoráveis que podem ser canalizados para tornar as pesquisas sobre o PLN não só mais "reais", isto é, mais próximas dos próprios pesquisadores, como mais integradas, capazes de reunir os diversos especialistas em equipes interdisciplinares.

O desenvolvimento de linguagens de programação exclusivas para a manipulação de símbolos, como as linguagens LISP e PROLOG, por outro lado, contribuiu para uma nova concepção de PLN. O computador, ou o *ordenador*, como os franceses e espanhóis preferem denominá-lo, embora tenha sido inicialmente explorado para realizar cálculos numéricos, passou também a ser concebido como uma máquina capaz de realizar um "trabalho geral de (re)ordenação" dos mais variados tipos de informação.[47] Com essa nova concepção, os computadores passaram a ser considerados máquinas universais de manipulação de cadeias de símbolos e, portanto, capazes de processar estruturas simbólicas complexas como, por exemplo, palavras, frases, árvores de representação sintática, e redes semânticas (GAZDAR & MELLISH, *op. cit.*).

A nova postura em relação à própria concepção de programação de computadores também desempenhou um papel decisivo para o salto qualitativo das pesquisas sobre o PLN. Sterling & Shapiro (1986), por exemplo, consideram a atividade de programação de

47 Como aponta Robert (1993.), o termo *ordenador*, cunhado pelos franceses na década de 70, reflete, de fato, essa outra concepção do papel dos computadores.

computadores parte do próprio processo de resolução de problemas: “a programação de computadores é concebida como uma atividade intelectualmente gratificante, uma ferramenta que nos auxilia na organização, na expressão, na experimentação e, até mesmo, na comunicação de nossas idéias”.

Esse encontro de fatores favoráveis foram responsáveis pela qualificação e redirecionamento da concepção dos estudos sobre o PLN, que deixam de ser considerados projetos lingüisticamente desmotivados e voltados apenas para os aspectos quantitativos das línguas naturais (*cf.* BOTT, 1976; BIDERMAN, 1978). Esses trabalhos, de fato, pouco dependiam dos lingüistas, uma vez que ficavam restritos a cálculos estatísticos, geralmente efetuados para a composição de listas de freqüência de ocorrências de palavras e de seus contextos de ocorrências ou para o estabelecimento de possibilidades combinatórias de itens lexicais.

Antes de encerrar este capítulo, duas observações são pertinentes.

A primeira diz respeito ao destaque que dou às discussões sobre os procedimentos de análise, reunidos sob a denominação genérica “compreensão das línguas naturais” (tradução do inglês *natural language understanding*). É de se esperar que um SPLN, com diferentes graus de sofisticação, seja capaz de simular os dois processos complementares de recepção e de produção de textos. Os trabalhos pesquisados, entretanto, têm enfatizado que a investigação dos procedimentos de recepção computacional de textos é tarefa prioritária. A justificativa para esse desequilíbrio, segundo Grishman (*op. cit.*: 159), é o fato de que, em geral, os SPLNs são projetados com finalidades

práticas, exigindo-se deles maior eficiência e precisão no processo de recepção do que no processo de produção de mensagens. Durante a fase de síntese, esses sistemas limitam-se simplesmente a construir seus "textos" de acordo com um conjunto fixo de padrões previamente armazenados na memória da máquina. Essa "simplificação", embora comprometa o "grau de naturalidade" do texto produzido, não chega a comprometer sua legibilidade. Durante a fase de análise, porém, um SPLN, para ser considerado minimamente eficiente, precisa estar preparado para desfazer uma série de ambigüidades e "interpretar" uma grande variedade de paráfrases de que o usuário se utiliza para se comunicar com a máquina. Por essa razão, as pesquisas sobre "geração de textos" (tradução do inglês *text generation*, empregada para denominar a outra sub-área de estudos do PLN) encontram-se ainda incipientes (*cf.* McKEOWN, *op. cit.*; APPELT, 1985; MYKOWIECKA, *op. cit.*). GRISHMAN (*op. cit.*: 159) chega a rotulá-las de "as primas pobres". **48**

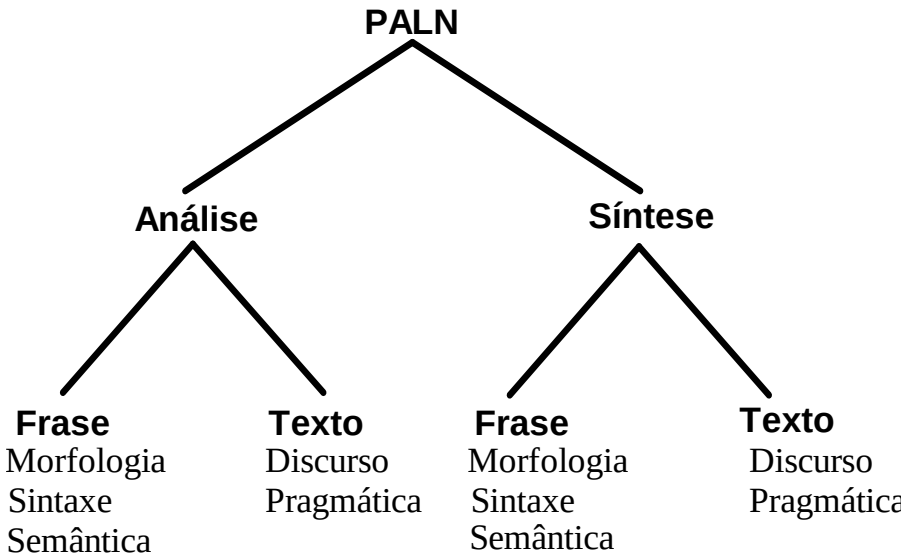
Ressalto, portanto, que o estudo do PLN, proposto nesta tese, focaliza exclusivamente o processamento automático de formas ortográficas dos enunciados lingüísticos, com ênfase nos procedimentos de análise.

---

48 Tomo o termo 'texto' para designar a manifestação lingüística do discurso, consistindo "em qualquer passagem, falada ou escrita, que forma um todo significativo, independente de sua extensão […] que se caracteriza por um conjunto de relações responsáveis pela tessitura do texto…" (FÁVERO & KOCH, *op. cit.*: 25); e o termo 'discurso' para designar qualquer "atividade comunicativa de um falante, numa situação de comunicação dada, englobando o conjunto de enunciados produzidos pelo locutor (ou por este e seu interlocutor, no caso do diálogo) e o evento de sua enunciação" (*op. cit. loc. cit*). Como características importantes, assumo que o texto constitui um objeto lingüístico que (i) possui uma extensão limitada, (ii) realiza alguma intenção comunicativa e (ii) sua interpretação resulta da capacidade do receptor de relacioná-lo ao universo do discurso, construído pelo receptor, a partir de um contexto situacional compartilhado (na situação de um diálogo, por exemplo) ou da reconstrução desse contexto, ou de parte dele.

A segunda refere-se à exclusão do tratamento computacional dos aspectos ligados aos estratos fonético e fonológico do sistema lingüístico, bem como às questões referentes aos elementos prosódicos, uma vez que fazem parte de um campo de estudos independente, porém, relacionado ao PLN. Reunidas sob a denominação "reconhecimento e produção da fala" (tradução do inglês "*speech recognition and production*"), essas pesquisas investigam as possibilidades de se projetarem sistemas computacionais capazes de reconhecer e interpretar frases e textos orais, concentrando esforços no equacionamento do problema da representação e reconhecimento dos sinais acústicos que compõem o fluxo da fala (*cf.* GEVARTER, *op. cit.*; BRISCOE, 1990), tarefa que apresenta características específicas e complexidades adicionais, justificando-se assim um tratamento à parte.

O esquema a seguir esuqematiza os domínios de estudo do PLN:



CAPÍTULO 3 – Uma estratégia de pesquisa para o PLN

> *"To build a knowledge system today, a knowledge engineer performs four types of functions: mining, molding, assembling, and refining. […] Knowledge, like a rare metal, lies dormant and impure, beneath the surface of consciousness. Once extracted, an element of knowledge must undergo several transformations before it can add value."*
>
> Frederik Hayes-Roth (1990: 294)

Aglutinação de esforços de disciplinas matrizes

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, encontram-se as lentes dos observadores direcionadas para determinados aspectos dos objetos lingüísticos, ocultando outros, que seriam apenas visíveis sob outras óticas.

Os lingüistas, num certo momento, por exemplo, ao observarem os fenômenos da linguagem com as "lentes do gerativismo", uma das correntes de pesquisa gramatical dominante desde seu nascimento na década de 50 (*cf.* HARLOW & VINCENT, *op. cit.*), focalizaram sua atenção até os limites da frase. Qualquer fenômeno a ela "transcendente" fugia do seu alcance. A fonologia e a morfologia foram minimizadas e a "grande e soberana" sintaxe ocupou o centro das discussões. A semântica, por sua vez, ficou praticamente posta de lado.

Afinal, como salienta Lyons (1979: 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 "lentes novas", ou "emprestadas" de estudiosos de outras disciplinas, as questões semânticas, e muitas outras novas questões, passaram a ser consideradas "lingüisticamente tratáveis": *análise do discurso* (*cf.* PRINCE, 1988), *pragmática* (*cf.* LEVINSON 1983; LEECH, *op. cit.*; HORN, 1988), *postulados conversacionais* (*cf.* KEMPSON, 1988), *atos de fala* (*cf.* SADOK, 1988), entre outras.

Esses fatos 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 "novas lentes" é, na maioria das vezes, fruto de influências de outras áreas do saber sobre a Lingüística.

Um exemplo bastante significativo dessas contribuições recíprocas pode ser encontrado em Chomsky (1957). Recorrendo à *teoria dos autômatos* (KORFHAGE, *op. cit.*), desenvolvida por matemáticos, Chomsky propôs seu modelo formal de análise gramatical que veio a revolucionar os estudos da linguagem. Como observa Lyons (1976: 63):

> "O passo revolucionário dado por Chomsky, no que diz respeito à lingüística, 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."

Eu acrescentaria que essa ousadia, além de ter trazido novo impulso para os estudos sintáticos, encontrou ressonância nos estudos matemáticos.

Ao estudar a possibilidade de criar modelos formais de gramática para as línguas naturais, com vistas à proposição da gramática gerativo-transformacional, Chomsky acabou também por inaugurar uma área de investigação essencial para os estudos computacionais: o estudo das *linguagens formais* (SUDKAMP, *op. cit.*) que, por sua vez, foi decisivo para a criação de linguagens de programação, compiladores e interpretadores, isto é, programas computacionais especializados que transformam uma determinada linguagem de programação em códigos executáveis pela máquina.. Como conseqüência, serviu também de estímulo para os estudos sobre o PLN, que até a década de 70 vinham sofrendo os efeitos negativos das experiências mal-sucedidas com a tradução automática.

Já 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, rigorosamente não deve ser considerada um desdobramento da Lingüística. Sua "lente", até meados da década de 60, centrava-se exclusivamente no estudo das linguagens formais e das linguagens de programação (*cf.* BALLARD & JONES, *op. cit.*: 133). Mesmo hoje, com o amadurecimento crescente dos estudos sobre o PLN, a Lingüística Computacional concentra-se em um único aspecto do empreendimento: o *estudo de algoritmos de análise morfológica* e *sintática* (EARLEY,

1970; KAY, 1985; HEARN *et al.*, *op. cit.*, KLAVANS, 1989). O estudo de sistemas de representação e os procedimentos computacionais de interpretação semântica e pragmático-discursivo ultrapassa seu domínio (*cf.* GRISHMAN, *op. cit.*; NIRENBURG *et al.*, *op. cit.*).

Com o estímulo proveniente da Lingüística, iniciado por Chomsky, e com influências diretas da Filosofia da Linguagem e da Psicologia, os estudos sobre o PLN passaram a abordar os mesmos temas dessas disciplinas matrizes: morfologia, sintaxe, semântica, pragmática, discurso, texto, aquisição da linguagem, entre outros (*cf.* CARBONELL & HAYES, 1990). Exemplos significativos que atestam as contribuições recíprocas que passaram a existir entre os estudos sobre o PLN e os estudos lingüísticos são: modelo do processamento de estruturas sintáticas (FRAZIER & FODOR, 1978; FODOR & FRAZIER, 1980); rede de transição ampliada, projetada para representar o processo de análise sintática, valendo-se da gramática gerativo-transformacional (WOODS, 1970); codificação e implementação de parcelas da gramática funcional proposta por Halliday (WINOGRAD, *op. cit.*);[49] modelo computacional dos atos de fala (COHEN & PERRAULT, 1979; ALLEN & PERRAULT, 1980); analisador sintático fundamentado na gramática gerativo-transformacional (MARCUS, *op. cit.*); teoria lingüística motivada por modelos computacionais (BRESNAN, 1982); estudo das propriedades matemáticas das línguas (PERRAULT; 1984); modelo computacional de geração de textos (McKEOWN, *op. cit.*; APPELT, 1985); modelo formal de interpretação semântica (DOWTY *et al.*, 1985); proposta de analisador sintático fundamentado na teoria chomskiana dos princípios e parâmetros (PRITCHETT, 1988 e 1989;

49 *Cf.* Halliday & Hasan (1976) e Halliday (1985).

PRITCHETT & REITANO (s/d)); modelo computacional da teoria da referência (KRONFELD, 1990); modelo computacional de interpretação semântica (HIRST, *op. cit.*).

Assim como a Língüística, a Inteligência Artificial também veio enriquecer os estudos sobre o PLN e, por meio deles, avançar 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 inferências* (REYTER, 1987; HOBBS *et al.*, 1990; CARPENTER & THOMASON, 1990);[50] o modelo de *redes semânticas*, criado para a representação da estrutura conceitual que serve de ancoragem para a estruturação do léxico (QUILLIAN, 1968) e as técnicas empregadas pela *engenharia do conhecimento* (HAYES-ROTH, *op. cit.*).

Mesmo que, historicamente, a Inteligência Artificial e a Lingüística Computacional, ambas consideradas ramificações da Ciência da Computação (*cf.* BALLARD & JONES, *op. cit.*; NIRENBURG *et al.*, *op. cit.*), tenham tomado para si o estudo do PLN, a sábia concepção de Winograd e os trabalhos mencionados, o colocam como um empreendimento interdisciplinar.[51] Dessa nova perspectiva, o PLN não

---

50 O campo denominado "representação do conhecimento", que estuda meios de criar sistemas formais de organização, representação e manipulação de informações, constitui uma das principais áreas de pesquisa da Inteligência Artificial (*cf.* BRACHMAN & LEVESQUE, *op. cit.*).

51 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

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, além de ser causa de significativas influências recíprocas.

Como mostra Petrick (1990), há 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, *op. cit.*).

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 mostrar que as gramáticas sintagmáticas livres de contexto eram inadequadas à caracterização da sintaxe das línguas naturais perderam sua força com a proposição das redes de transição ampliadas feita por 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 (*op. cit.*) mostra que as gramáticas gerativas, com seu grande número de dispositivos formais, são completamente inadequadas a servir de modelo de processamento das estruturas lingüísticas pelos falantes. A partir dessa análise, constrói um novo modelo de gramática, sem as

---

Shannon, Oliver Selfridge, Nathaniel Rochester, entre outros. A disciplina *Lingüística Computacional*, por sua vez, cujo nome foi cunhado em 1967 por David Hays (*cf.* MORENO FERNÁNDEZ, *op. cit.*: 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.

"transformações" – a gramática sintagmática generalizada. 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 suas "lentes" a linguagem humana, suas funções, sua manifestação nas diferentes línguas, sua estrutura interna, sua relação com a realidade, com os processos de raciocínio e com o comportamento verbal.

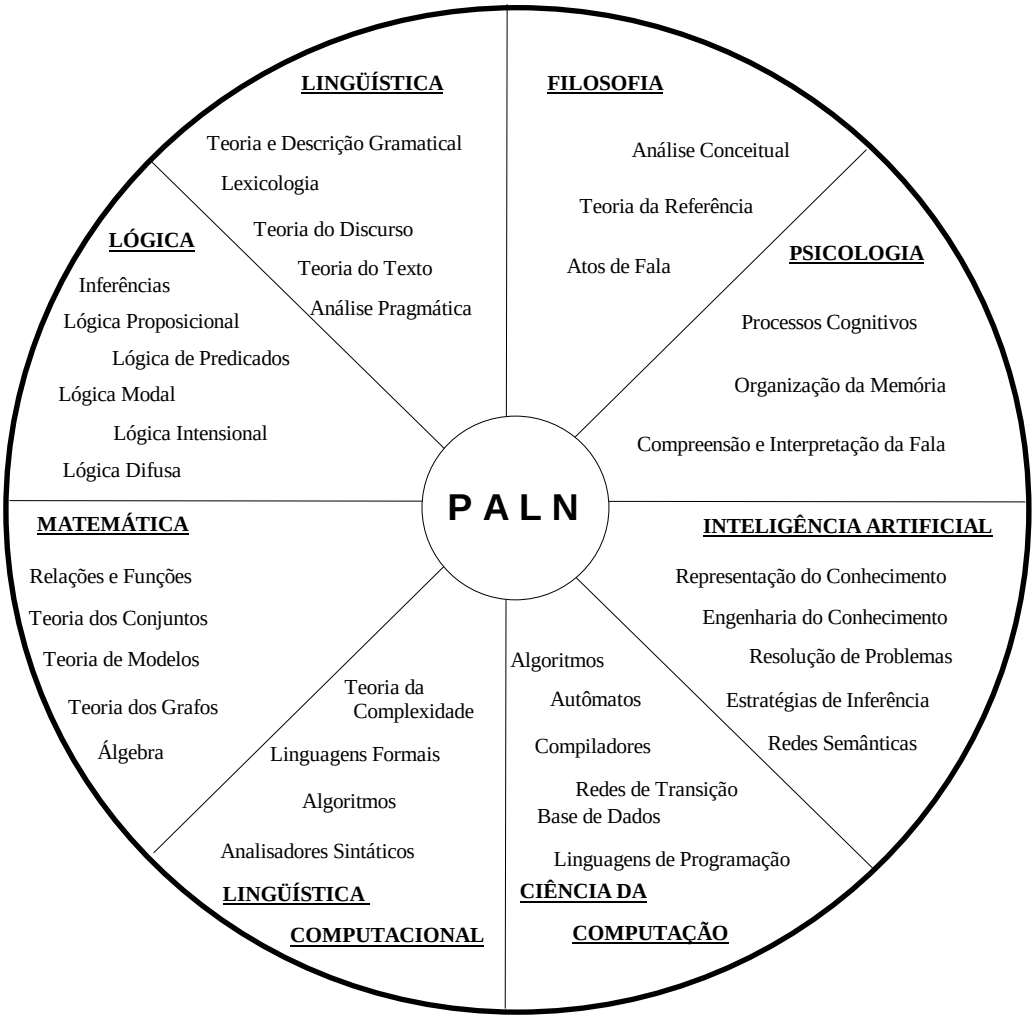
Sobrepõe-se também a domínios da lógica, matemática, ciência da computação, lingüística computacional e inteligência artificial, ao procurar, nestes instrumentos, estratégias indispensáveis à construção e à implementação dos modelos. São exemplos: sistemas de representações formais, como o cálculo de 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 sintática; os sistemas de representação do conhecimento de mundo, de crenças; as estratégias de resolução de problemas e de organização da informação.

Há que se observar também que existem sobreposições entre a própria lingüística, de um lado, e a filosofia, a lógica, a matemática, a ciência da computação e a inteligência artificial, de outro. A mais clássica delas é a sobreposição que se constata entre os estudos da linguagem, filosofia e lógica (*cf.* FREGE, 1990; AUSTIN, 1962 e 1990; GRICE, 1990; SEARLE, 1990a e 1990b; REICHENBACH, 1947; LEHMANN *et al.*, 1985; BARWISE & PERRY, 1990). Lembre-se de que os estudos da linguagem originaram-se no seio da filosofia e da

lógica: 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).

Importantes contribuições mútuas que ocorreram neste século merecem destaque (*cf.* 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. Sob 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 de 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. Já a matemática, a ciência da computação e a inteligência artificial só muito recentemente têm instrumentalizado a lingüística no equacionamento de alguns de 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*, provenientes dos estudos sobre a inteligência artificial são empregados por teorias semânticas e teorias do discurso.

No esquema abaixo, classificados segundo as disciplinas matrizes, apresento a sistematização dos principais recursos teóricos e metodológicos de que o estudo do PLN dispõe:



Logo, o campo de estudos sobre o PLN não poderia deixar de ser um domínio de pesquisas privilegiado, amplo e fecundo, uma vez que a construção do corpo de conhecimentos necessários para a implementação de sistemas computacionais com esse grau de sofisticação necessariamente exige a seleção, a organização, a representação e a codificação de uma variedade de informações na

Dias da Silva, B.C. (1996)

complexa tarefa de criar um simulacro computacional da “competência” e do “desempenho” lingüísticos.

Nesse domínio, pesquisas interdisciplinares poderão encontrar um solo f'értil para germinar.

De fato, Sanders & Sanders (1989: 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, e 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 desenvolvimento de SPLNs são condições essenciais para a sua solidificação. Desenvolver pesquisas integradas e lingüisticamente fundamentadas sobre o PLN pode deixar de ser uma utopia.

Estratégia de pesquisa para o 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 desenvolvimento de SPLNs. Nesse sentido, proponho uma

estratégia de pesquisa para o estudo do PLN que envolve equacionar questões em três domínios: **lingüístico**, **representacional** e **implementacional**. A proposta, decorrência de muitas reflexões, encontra sua motivação e fundamentação nos projetos de desenvolvimento de sistemas computacionais que visam à construção de *bases de conhecimentos* (HAYES-ROTH, *op. cit.*) e na *programação automática* (BIERMANN, *op. cit.*). Esses sistemas são projetados para representar complexos de conhecimentos e aplicá-los automaticamente no processo de *resolução de problemas* (AMAREL, *op. cit.*).[52]

O processo de construção de sistemas especializados dessa natureza pressupõe a especificação dos tipos de conhecimentos que os especialistas possuem, como esse conhecimento é armazenado, acessado, aplicado e adquirido (*cf* SCHANK & RIESBECK, *op. cit.*: 2). Assim, projetar um sistema de computador que simule parcelas da competência e atuação de um sintaticista, por exemplo, pressupõe a especificação de conhecimentos e habilidades que um especialista dessa área possui. De modo análogo, projetar um sistema de computador que simule parcelas da competência e do desempenho lingüísticos humanos, pressupõe a especificação de conhecimentos e habilidades que os falantes, especialistas nesse domínio, possuem.

Assumindo a concepção de PLN de Winograd, verificamos que, para simular uma língua natural de modo satisfatório, um SPLN precisa conter vários sistemas de "conhecimentos" e realizar uma série de atividades "cognitivas":

- possuir um "modelo simples de sua própria mentalidade";

---

52 Os diversos componentes serão apresentados no sexto capítulo, quando será proposta uma arquitetura para um SPLN.

- 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.

O termo "conhecimento" (do inglês *knowledge*) é um "termo guarda-chuva" ("*blanket term*"), empregado para denotar qualquer tipo de informação manipulada por um sistema computacional (*cf.* NIRENBURG *et al.*, *op. cit.*: 219). Seguindo essa prática, os pesquisadores em inteligência artificial costumam dizer que os SPLNs "possuem" vários tipos de conhecimentos e "sua mentalidade" permite utilizar estratégias de inferência. Como já observara no primeiro capítulo, a antropomorfização da máquina é uma constante e, muitas vezes, inevitável. Procurando evitar discussões controvertidas sobre possibilidades de criação de uma inteligência artificial, passo a empregar os termos informação e mecanismos de inferência, respectivamente.

A analogia que estou construindo permite conceber os SPLNs 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 lingüísticos. Assim, o estudo do PLN pode ser concebido como um tipo de "engenharia do conhecimento lingüístico" e beneficiar-se da

estratégia desenvolvida para o campo denominado "engenharia do conhecimento" (*cf.* HAYES-ROTH, *op. cit.*).

De modo semelhante ao processo de construção de um "sistema de conhecimento" (do inglês *knowledge system*), a montagem de SPLNs exige o desenvolvimento de, no mínimo, três etapas: "extração do solo" (explicitação dos conhecimentos e habilidades lingüísticas), "lapidação" (representação formal desses conhecimentos e habilidades) e "incrustação" (o programa de computador que codifica essa representação).[53] O esquema a seguir sintetiza as tarefas previstas e especifica os resultados esperados de cada etapa:

Tarefas	**Resultados**
• explicitar o conhecimento	• representação lingüística
• representá-lo formalmente	• representação computacional
• codificá-lo e implementá-lo	• SPLN

Os *estudos da linguagem desenvolvidos pela ciência cognitiva* também propõem três níveis de abordagem do processamento humano da linguagem que, grosso modo, 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

53 O autor emprega esses termos intencionalmente e justifica a escolha, por considerar que o conhecimento necessário para montar um sistema especializado dessa natureza, como um minério bruto, precisa primeiro passar por essas etapas para depois ser utilizado.

> 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 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."[54]

A partir dessas considerações de Lasnik é possível montar as seguintes correlações:

Ciência cognitiva	**Objeto da análise**
• nível lingüístico	• conhecimento lingüístico
• nível representacional	• representação computacional
• nível implementacional	• suporte neurológico da linguagem

Por fim, Barton, Berwick & Ristad (*op. cit.*: 96-7) esclarecem que a teoria da competência lingüística, pertencente ao "nível computacional", deve explicar quais são as estruturas calculadas e por que, 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, 3) e as considerações sobre a "competência pragmática", chegamos ao esquema, a seguir, que resume a estratégia para o equacionamento do PLN:[55]

---

54 Grifo meu. O nível representacional corresponde ao que estou denominando computacional e este corresponde ao que estou denominando lingüístico.

55 Chomsky coloca duas questões programáticas para os estudos da linguagem: (i) Em que consiste o conhecimento lingüístico? (Teoria da Competência); (ii) Como esse conhecimento é colocado em uso? (Teoria do Desempenho). Observo também que uma teoria do desempenho deverá ser entendida como uma teoria da gramática acoplada a uma teoria que caracterize os mecanismos de processamento da linguagem no processo de produção e

DOMÍNIOS	PROBLEMAS	RECURSOS
Lingüístico	explicitar o conhecimento e o uso lingüístico	teorias da competência e do desempenho
↓↑	↓↑	↓↑
Representacional	representá-los	linguagens formais de representação
↓↑	↓↑	↓↑
Implementacional	coficar as representações	linguagens de programação e sistemas computacionais

Como ressalta Halvorsen (*op. cit.*: 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 uso lingüísticos envolve questões do domínio lingüístico, uma vez que é nessa fase que os fatos da língua e do seu uso são especificados. Conceitos, termos, regras,

compreensão de frases. Em outras palavras, o uso pode ser entendido como um “processador lingüístico”. É importante notar que fenômenos lingüísticos complexos, como é o caso do *objeto nulo*, por exemplo, podem ser explicados em termos da interação entre a gramática (competência) e o processador (desempenho).

princípios, estratégias de resolução de problemas e formalismos lingüísticos são os elementos trabalhados. No domínio da representação, questões referentes à escolha ou à proposição de sistemas de representação, que incluem, por exemplo, a lógica, redes semânticas, regras de reescrita e *frames*, bem como estratégias de codificação dos elementos trabalhados no domínio anterior, entram em foco. 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.

Fases de construção de SPLNs

Os três domínios acima delimitados, por sua vez, podem ser reinterpretados como três fases sucessivas do desenvolvimento de um SPLN particular, ou parte dele:

- **Fase Lingüística**: 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.

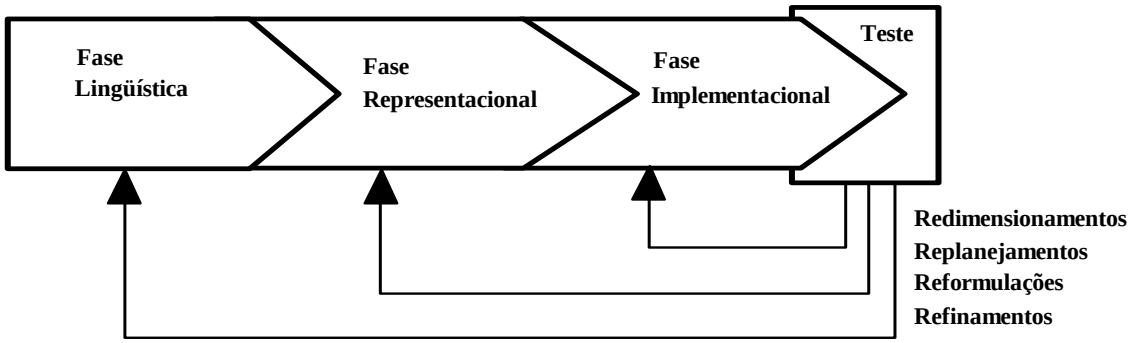
- **Fase Representacional**: construção conceitual do sistema, envolvendo a seleção e/ou proposição de sistemas formais de representação para os resultados propostos pela fase anterior. Nesta fase, projetam-se as representações lingüísticas e extralingüísticas em sistemas formais computacionalmente tratáveis.

- **Fase Implementacional**: codificação das representações elaboradas durante a fase anterior em termos de linguagens de programação e planejamento global do sistema. Nesta fase, além de transformar as representações da fase anterior em programas computacionais, estudam-se as questões referentes à integração conceitual e física dos vários componentes envolvidos, bem como questões referentes ao ambiente computacional em que o sistema será desenvolvido e implementado.

Proponho que as três fases sejam desenvolvidas sucessiva, progressiva e ciclicamente: as representações parciais resultantes das duas primeiras fases podem ser implementadas e, finalmente, testadas, completando, assim, um ciclo.[56] Dessa forma, testes de adequação e de desempenho poderão contribuir para o aprimoramento dos resultados

56 Essa decomposição modular de um problema complexo em subproblemas espelha também uma estratégia de trabalho bastante difundida e profíqua nos estudos computacionais (*cf*. RICH, 1983; WINSTON, 1984). Bresnan (1982), no âmbito dos estudos lingüísticos, emprega estratégia semelhante na construção da própria teoria léxico-funcional.

alcançados em cada fase. A dinâmica do processo pode ser assim visualizada:



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

Logo, partindo-se de uma descrição informal, chega-se a uma representação interpretável pela máquina. Uma simples regra sintática, por exemplo, seria assim desenvolvida:

FASE LINGÜISTICA	**DESCRIÇÃO INFORMAL** "Uma frase pode ser composta pela concatenação de um sintagma nominal e de 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 de pessoa. O caso gramatical do sujeito é nominativo e o verbo encontra-se na forma finita." (QUIRK & GREENBAUM, 1973) **FORMALISMO LINGÜÍSTICO** **F → SN SV** (↑ SUJEITO) = ↓ ↑ = ↓ (KAPLAN & BRESNAN, 1982)

FASE REPRESENTACIONAL	**REPRESENTAÇÃO COMPUTACIONAL** **Regra Sintática:** F → SN SV **Expecificações:** <SN pessoa> = <SV pessoa> <SN número> = <SV número> <SN caso> = nominativo <SV forma verbal> = finita (SHIEBER, 1986)

	IMPLEMENTAÇÃO EM PROLOG
FASE IMPLEMENTACIONAL	**f(P0,P):-sn(Pessoa,Número,Caso, P0,P1),** **sv(Pessoa,Número,Caso,P1,P).** (CLOCKSIN & MELLISH, 1981)

A Fase Representacional, entre a representação abstrata fornecida pela teoria lingüística e o programa de computador, além de ser necessária, é estrategicamente positiva por duas razões: (i) como um projeto arquitetônico, as representações formais contêm todas as informações necessárias para a construção do SPLN e, por princípio, (ii) não estão diretamente "comprometidas" com nenhuma linguagem de programação específica, o que garante maior "transportabilidade" dos resultados já alcançados para tipos de máquinas diferentes, empregando linguagens de programação e algoritmos também diferentes.

CAPÍTULO 4 – Equacionamento do domínio lingüístico

> *"A language comprehension program must have considerable knowledge about the structure of language itself, including what the words are, how to combine the words into sentences,…. However, a program cannot completely simulate linguistic behavior without first taking into account an important aspect of what makes humans intelligent – their general world knowledge and their reasoning ability."*
>
> James Allen (1987: 9)

O Homem não só é capaz de criar e compreender os mais variados tipos de textos orais e escritos codificados nas diversas línguas espalhadas pelo mundo como também é capaz de realizar outras tantas atividades de natureza meta-lingüística que abrangem desde a identificação e interpretação de um simples morfema até a tradução de textos de uma língua para outra. A habilidade, naturalidade e eficiência com que desempenha o gerenciamento desses processos, envolvendo a produção, manipulação e recepção de uma quantidade massiva e variada de objetos lingüísticos são surpreendentes. Como salienta Garrett (1990: 133):

> "Todos os dias, todo ser humano normal emite milhares de palavras sob a forma de frases e segmentos conversacionais e ouve o dobro desse número ou, até mesmo, mais. Cada palavra emitida precisa ser identificada dentro de um conjunto de 50 mil, ou mais, formas em menos de um terço de segundo e integrada em uma estrutura

> que expresse corretamente o significado pretendido pelo falante. Além disso, os processos que resultam nessa associação de significado e forma do enunciado são, em geral, desencadeados sem a atenção, ou intenção, consciente do locutor ou do interlocutor."

Já Schank & Riesbeck (*op. cit.*: 3) atentam para a natureza explícita e rígida dos objetos lógicos e para a subespecificação inerente e estratégica dos objetos lingüísticos:

> "As línguas naturais não são como as linguagens de programação. Enquanto estas são criadas segundo as exigências de completude e exatidão impostas pela lógica, aquelas desenvolveram-se para preencher as necessidades comunicativas verbais humanas. Quando sabemos o assunto que está sendo discutido, umas poucas pistas verbais são suficientes para nos orientar a compreensão de textos. Essa característica humana, isto é, o fato de sermos especialistas na tarefa de suprir informações implícitas em praticamente tudo o que lemos ou ouvimos, permite que as línguas naturais deixem de explicitá-las exaustivamente. Essa subespecificação, característica das formas lingüísticas, é marcadamente evidenciada pela profusão de enunciados contendo elipses, referências anafóricas e palavras extremamente ambíguas."

A complexidade lingüística

Considere, por exemplo, a atividade que consiste em ler um texto escrito e, em seguida, responder a perguntas sobre ele:[57]

[57] Ao longo da análise, introduzirei fenômenos adicionais que não estão presentes no exemplo, mas que são relevantes para a exposição dos problemas.

Joaquim Cruz: “Me livrei de um sonho ruim: nele, eu não sabia mais o horário da prova; perguntava a um colega, que também não sabia, e chegava atrasado na pista.” **58**

1. Por que o atleta diz que o sonho era ruim?
2. O que o atleta pergunta ao colega?
3. Por que, no sonho, o atleta chega atrasado na pista?
4. O atleta perde a competição no mundo do sonho?

Um SPLN capaz de responder a essas poucas perguntas e, assim, mostrar que, de certa forma, “compreendeu” o pesadelo de Joaquim Cruz, precisa incorporar uma quantidade considerável de “conhecimentos” e “habilidades”.

De início, o sistema precisa conter informações sobre o próprio *layout* do texto impresso. Essas informações explícitas como, por exemplo, os espaços em branco que separam as palavras e os sinais de pontuação específicos e que demarcam os limites de uma frase ou período, são essenciais e facilitam significativamente o processo de segmentação das unidades lingüísticas.**59** Línguas como o turco, cuja representação escrita não contém espaços em branco entre as palavras, colocam uma complexidade adicional aos projetistas de PLN (*cf.* GAZDAR & MELLISH, *op. cit.*: 144).

Além dessas informações, o sistema precisa também “conhecer”, pelo menos de maneira parcial, a língua portuguesa e como

58 (VEJA, 22/07/92: p.9)

59 Há também outros sinais de pontuação específicos e outras convenções ortográficas e tipográficas, isto é, recursos empregados para representar graficamente uma multiplicidade de informações lingüisticamente relevantes, como pausas, ênfases, termos técnicos, palavras estrangeiras e modalizações.

ela é colocada em uso pelos falantes. Assim, o sistema precisa "saber", por exemplo, que:

- seqüências de símbolos como N E L E são resultantes de convenções meramente ortográficas: EM ELE
- seqüências de símbolos como L I V R E I são resultantes de processos flexionais da língua (*cf.* BAKER, 1988): LIVR + EI;
- seqüências como *-ei*, *-ia* e *-va* são morfemas flexionais, e, portanto, elementos portadores de informações gramaticais: NÚMERO, PESSOA, TEMPO, MODO, VOZ e ASPECTO; e seqüências como *-a*, em *atrasada*, são indicadores de GÊNERO;
- seqüências como COMPETIÇÃO são resultantes de processos derivacionais: COMPETIR (Verbo) → COMPETIÇÃO (Substantivo)
- as seqüências atômicas pertencem a uma determinada categoria gramatical: Substantivo (*sonho*), Verbo (*livr-*), Preposição (*em*), Adjetivo (*ruim*) e Advérbio (*mais*), Pronome (*me*), Determinante (*um*), Complementador (*que*), Conjunção (*e*);
- as categorias nucleares sistematicamente projetam constituintes mais complexos (*cf.* JACKENDOFF, 1977): Determinante + Substantivo + Adjetivo = Sintagma Nominal (*um sonho ruim*); Preposição + Sintagma Nominal = Sintagma Preposicional (*de um sonho ruim*); Verbo + Sintagma Preposicional = Sintagma Verbal (*me livrei de um sonho ruim*); Sintagma Nominal + Sintagma Verbal = Frase (*Ø Me livrei de um sonho ruim*); Frase + Frase = Frase (*perguntava a um colega + que também não sabia*); Frase + Conjunção + Frase = Período

(*perguntava a um colega que também não sabia e chegava atrasado na pista*);

- os constituintes que integram a frase desempenham funções gramaticais específicas (*cf.* BRESNAN, 1981): SUJEITO (*eu*), PREDICADO (*perguntava a um colega*), OBJETO INDIRETO (*a um colega*);
- o valor dos traços de NÚMERO e PESSOA do sujeito e do verbo precisam coincidir: *eu* [PESSOA(1), NÚMERO(sg)] = *livrei* [PESSOA(1), NÚMERO (sg)];
- existem muitos elementos gramaticalmente opcionais agregados às frases: os modificadores nominais *ruim*, *da prova* e os modificadores verbais *nele*, *mais*, *também*, *não*, *atrasado*;
- existem elementos "apagados", mas que são gramaticalmente recuperáveis: o objeto do verbo *perguntar*, o objeto da segunda ocorrência do verbo *saber* e os sujeitos "ocultos";
- existem elementos co-referenciais: o pronome *me* e o sujeito do verbo *livrar-se de*, *um colega* e o sujeito da segunda ocorrência do verbo *saber*;
- o constituinte *O que* na pergunta 2 foi deslocado de seu local de origem (*cf.* CINQUE, 1990; MANZINI, 1992): *O que*$_{i}$ *o atleta perguntou* v_{i} *ao colega*?
- os predicadores possuem uma estrutura de argumentos temáticos (*cf.* GRUBER, 1965; FILLMORE, 1968 e 1977; JACKENDOFF, 1972, 1983 e 1990; SCHANK & ABELSON, 1977; SCHANK, 1982; SCHANK & RIESBECK, *op. cit.*; BRUCE & MOSER, 1990; BORBA, 1991; GRIMSHAW, 1992): *livrar-se de*

(BENEFICIÁRIO,TEMA), *saber* (EXPERIMENTADOR,TEMA), *perguntar* (AGENTE,TEMA,META); *chegava* (AGENTE,LOCATIVO);

- os papéis temáticos dos predicadores associam-se a funções gramaticais específicas (*cf.* LEVIN, 1987): *livrar-se de* (BENEFICIÁRIO = SUJEITO, TEMA = OBJETO);
- um mesmo verbo pode projetar estruturas sintáticas diferentes: *Joaquim afundou o barco com uma bomba*, *A bomba afundou o barco*, *O barco afundou*;
- as frases do texto são declarativas;
- as frases 1, 2, 3 e 4 sobre o texto são formas interrogativas;
- as formas interrogativas solicitam informações que estão codificadas no próprio texto ou que podem ser inferidas a partir dele;
- seqüências como *de* em *livrei de*, *um* em *um sonho*, e *a* em *a pista*, por exemplo, são formas que desempenham funções gramaticais, enquanto que seqüências como *sonho* e *pista*, por exemplo, são elementos que possuem uma intensão e podem ser empregados extensionalmente;
- as expressões *sonho ruim* e *pesadelo* são expressões lingüísticas semanticamente equivalentes;
- existem classes de tipos e subtipos semânticos (T) e restrições selecionais (R) que regulam os julgamentos sobre anomalias semânticas (A) (*cf.* KATZ, 1972; ALLEN, *op. cit.*): (T) *Joaquim Cruz tem um imóvel* (casa) *e o colega tem dois* (apartamentos); (A) *Joaquim Cruz encontrou duas pistas* (lugar onde se pratica esportes) *e o colega três* (vestígios); (A) *Joaquim Cruz perguntava a um colega*

[+HUMANO] *e à pista* [-HUMANO]; (R) *a maçã* [+OBJETO FÍSICO] *verde* [COR]; (R) *a maçã* [+OBJETO FÍSICO, +FRUTA] *verde* [-MADURO]; (A) *a idéia* [-OBJETO FÍSICO] *verde* [+COR]; (R) *a idéia* [-OBJETO FÍSICO] *verde* [+ECOLÓGICO];

- existem modulações e seleções contextuais que destacam, ofuscam ou transferem certos traços semânticos (*cf.* CRUSE, 1986; PUSTEJOVSKY BOGURAEV, 1991): embora *manteiga* seja conicamente [+SÓLIDO], o efeito contextual provocado pelo verbo *despejar* em *O atleta despejou a manteiga na pista* transformou o traço [+SÓLIDO] em [-SÓLIDO], além de acrescentar o traço [+QUENTE], evocando assim outro aspecto do referente associado à expressão *a manteiga*; em *O atleta precisa consertar o equipamento* e *O atleta precisa lavar o equipamento*, dois aspectos distintos do equipamento; fato semelhante ocorre com a expressão *pista* em *pista rápida* (aspecto funcional), *pista esburacada* (aspecto formal), *pista sem acostamento* (aspecto constitutivo), *pista mal feita* (aspecto "genético");

- a interpretação do anafórico nulo Ø contido na frase *perguntava Ø a um colega* depende da frase *eu não sabia mais o horário da prova*;

- *ele* (em *nele)* e *um sonho* referem-se à mesma entidade (*cf.* BOSCH, 1983);

- existem verbos como achar e querer que funcionam como operadores referencialmente opacos: a partir de *Pedro é o melhor corredor* e *Joaquim Cruz acha que Pedro venceu a prova* não se pode concluir que *Joaquim Cruz acha que o melhor corredor ganhou a prova*; porém, a partir de *Pedro é o melhor corredor* e *Pedro venceu a prova* é possível concluir que *O melhor corredor venceu a prova*;

- o referente dos pronomes *me*, *eu* e o sujeito dos verbos *livrar-se de*, *perguntar*, *chegar* e o sujeito da primeira ocorrência do verbo *saber* é o produtor do enunciado;
- os "mundos específicos" em que os eventos ocorrem precisam ser conhecidos. E preciso distinguir entre o "mundo do sonho" e o "mundo real": Joaquim Cruz deixa de ter um pesadelo no "mundo real" em um "tempo do mundo real" anterior ao momento em que relata o fato. No "mundo do sonho", ocorre uma seqüência de eventos: Joaquim Cruz esquece o "horário da prova do sonho", pergunta esse horário a um "colega do sonho" e perde "a competição do mundo do sonho";
- o *foco de atenção* (*cf.* GROSZ & SIDNER, *op. cit.*) do produtor do texto é o *horário da prova*;
- uma série de informações extralingüísticas são necessárias para "decidir" sobre as questões relacionadas ao "mundo dos humanos". Esse conhecimento inclui, por exemplo, saber a respeito dos "objetivos humanos típicos", do "mundo onírico" e do "mundo dos esportes e competições".

Além de possuir todas essas informações complexas, o sistema precisa ainda enfrentar o **problema da ambigüidade** das formas lingüísticas, seja ela **local** ou **global**.

A ambigüidade é **global**, quando toda a seqüência de palavras, que compõem a frase, projeta mais de uma estrutura oracional gramaticalmente bem-formada potencial. A frase do inglês *John saw the woman in the park with a telescope* tornou-se clássica como um exemplo

desse tipo de ambigüidade. Há pelo menos quatro interpretações possíveis para essa frase:

- [$_{F}$ [$_{SN}$ John] [$_{SV}$ saw [$_{SN}$ the woman] [$_{SP}$ in the park] [$_{SP}$ with a telescope]]]

  "João estava no parque e viu a mulher através de um telescópio"

- [$_{F}$ [$_{SN}$ John] [$_{SV}$ saw [$_{SN}$ the woman [$_{SP}$ with a telescope]] [$_{SP}$ in the park]]]

  "João estava no parque e viu que a mulher tinha um telescópio"

- [$_{F}$ [$_{SN}$ John] [$_{SV}$ saw [$_{SN}$ the woman [$_{SP}$ in the park]][$_{SP}$ with a telescope]]]

  "João viu a mulher que estava no parque através de um telescópio"

- [$_{F}$ [$_{SN}$ John] [$_{SV}$ saw [$_{SN}$ the woman [$_{SP}$ in the park][$_{SP}$ with a telescope]]]]

  "João viu que a mulher que estava no parque tinha um telescópio"

Outros exemplos desse tipo de ambigüidade são ilustrados por frases como: *A porta perto da entrada com a placa ''Só Convidados'' estava trancada* e *Pedro vendeu o carro que Maria comprou com sacrifício*.

Além desse tipo de ambigüidade estrutural, há ainda as ambigüidades lexicais, temáticas, referenciais e "pragmáticas".

As ambigüidades lexicais podem ocorrer devido a três fenômenos: polissemia, homonímia e categorização gramatical. Tanto a polissemia quanto a homonímia são fenômenos observados no interior de uma mesma categoria sintática. Uma palavra polissêmica apresenta um conjunto de significados relacionados: o verbo *abrir*, por exemplo, pode significar 'desdobrar, expandir, revelar, iniciar, separar, descerrar, criar

aberturas', e assim por diante. Já as palavras homônimas são palavras que possuem a mesma forma com significados totalmente diferentes: *manga* (fruta e parte de uma peça de vestuário), *banco* (instituição financeira e local para se sentar) e *canto* (ângulo e som musical).

As ambigüidades categoriais referem-se ao fato de uma mesma forma lexical pertencer a classes sintáticas distintas: *cara* (adjetivo e substantivo), *prova* (substantivo e verbo), *canto* (verbo e preposição), entre outras.

Como as preposições comumente sinalizam papéis temáticos, as ambigüidades temáticas ocorrem quando uma mesma preposição sinaliza funções temáticas diferentes. Por exemplo, na frase *Maria trouxe um carro para Pedro*, a preposição *para* pode introduzir o DESTINATÁRIO ou o BENEFICIÁRIO da ação. Nesse caso, teríamos uma espécie de "homonímia temática": *para* (DESTINATÁRIO e BENEFICIÁRIO). Neste ponto, é também importante notar que as preposições nem sempre marcam funções temáticas. Com efeito, os sintagmas preposicionais podem modificar verbos, adjetivos, substantivos, outros sintagmas preposicionais e advérbios (LEMLE, *op. cit.*: 170). Em *Maria vendeu a casa de jogos com desconfiança*, o sintagma preposicional *de jogos* modifica *casa*, e o sintagma preposicional *com desconfiança* qualifica o verbo; em ambos os casos não se trata de funções temáticas.

Determinar os referentes de elementos pronominais é um quinto foco de ambigüidades. Na frase, *Coloquei o pão sobre o balcão e o comi*, tanto o referente de *o pão* quanto o referente de *o balcão* estão sintaticamente "autorizados" para "preencher" o valor do pronome *o*.

A ausência de correspondência um-a-um entre forma gramatical e função comunicativa de uma expressão lingüística é a fonte das ambigüidades pragmáticas. Considere, por exemplo, dois expedientes sintáticos do português: omissão do sujeito *você* nas orações imperativas (1) e as interrogativas parciais (2):

(1) Copie.

(2) Quem você conheceu?

Cada uma dessas formas pode preencher funções retóricas diversas: [a] fazer um pedido, [b] ameaçar, [c] reclamar, [d] solicitar informação e [e] expressar surpresa:

[a] *Copie, por favor.*

[b] *Copie, que eu te dou zero.*

[c] *Copie! É só isso que sabe falar!*

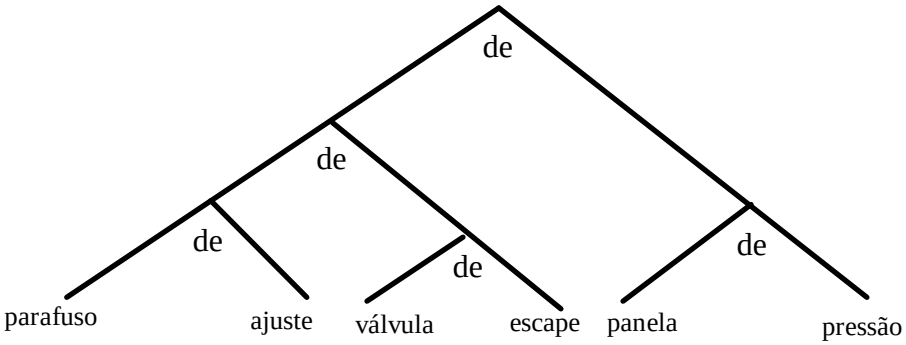
[d] *Quem você conheceu ?*

[c] *Quem você conheceu !?*

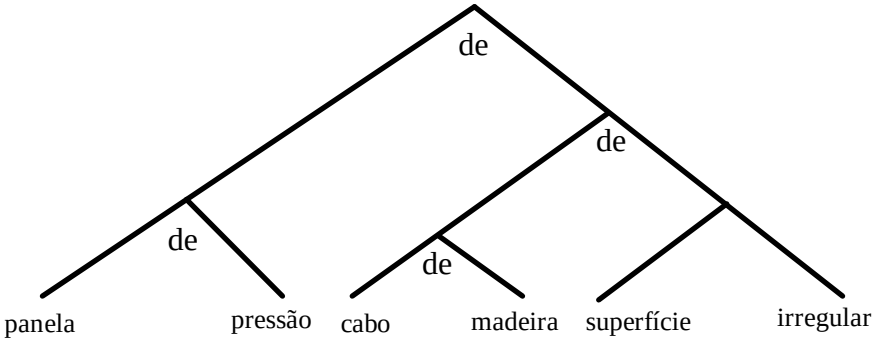
Já as ambigüidades são **locais** quando apenas partes da seqüência de palavras que integram a frase projetam estruturas gramaticalmente bem-formadas. Por exemplo, a frase *A empresa que comprou a Universal vendeu a Borland* poderia ser erroneamente analisada pela máquina como *A Universal vendeu a Borland*.

A construção de sintagmas nominais complexos, contendo sintagmas preposicionais, são também complicações adicionais. Vejamos alguns exemplos:

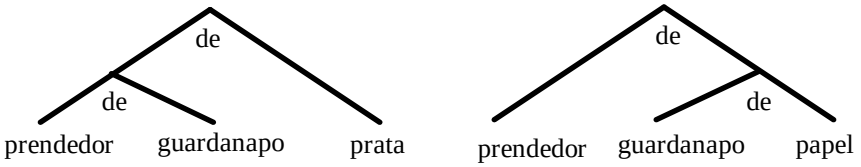
parafuso de ajuste de válvula de escape de panela de pressão



panela de pressão de cabo de madeira de superfície irregular



prendedor de guardanapo de prata *prendedor de guardanapo de papel*



Por fim, há frases, conhecidas como g*arden path sentences* (*cf*. PRITCHETT, 1988), que fatalmente levam o leitor a atribuir lhes uma estrutura sintática incorreta, obrigando-o a reprocessá-la para encontrar a estrutura correta. A ambigüidade entre uma oração imperativa e uma interrogativa só é resolvida, quando lemos a oitava palavra de cada uma das frases a seguir:

Have the students who missed the exam take it today.
Have the students who missed the exam taken it today ?

Não fosse a pausa na modalidade oral, ou a vírgula na modalidade escrita, a frase abaixo apresentaria problema semelhante:

Enquanto Paula estava ocupada tricotando seu casaco de lã estava secando.

Línguas como o português, que não apresentam a inversão sujeito-verbo auxiliar nas interrogativas, mecanismo típico do inglês, apresentam um problema adicional, porque, enquanto não se detectar o sinal gráfico de interrogação, não se pode decidir se a frase é ou não uma interrogativa. O recurso gráfico de se colocar um sinal de interrogação no início e outro no final de uma frase interrogativa, típico dos textos em espanhol, seria uma alternativa plausível para solucionar esse problema.

Ao se construir um SPLN não se pode perder de vista que, mesmo contando com condições de boa-formação, reponsáveis pela eliminação de objetos lingüísticos mal-formados, o fenômeno da ambigüidade das formas e das funções lingüísticas manifesta-se em todos os níveis de análise:

Nível	Condições de boa-formação	Tipo de Ambigüidade
Morfológico	regras de flexão e derivação	analítica (identificação e delimitação dos morfemas)
Sintático	regras da gramática	estrutural e categorial
Semântico	restrições selecionais	temática, semântica e quantificacional
Pragmático-Discursivo	princípios conversacionais	funcional (correspondência não biunívoca entre as funções pragmáticas e suas realizações)

Uma teoria lingüística

Diante da complexidade da linguagem, muitos modelos de análise têm sido propostos. Há muito, acredita-se que deve haver uma arquitetura universal subjacente a todas as línguas. Apesar da enorme variabilidade das formas de expressão encontradas nas várias línguas do mundo, postula-se a existência de uma gramática universal, um conhecimento específico e comum a todos os indivíduos da espécie humana. Em termos bastante genéricos, a teoria lingüística focaliza, de um lado, a caracterização dessa faculdade da linguagem e, de outro, como essa competência lingüística é colocada em uso pelos falantes de uma língua.

Noam Chomsky talvez possa ser nomeado, atualmente, o principal representante dessa concepção universalista dos fenômenos da linguagem. Pode-se mesmo afirmar que, entre suas grandes contribuições, uma das mais importantes tenha sido a proposição de uma teoria lingüística capaz de acoplar as idéias universalistas sobre a linguagem a um modelo formal de análise lingüística (*cf.* McCLOSKEY, 1989). Sua influência tem sido decisiva na concepção e no desenvolvimento da teoria lingüística desde a publicação de seu trabalho

Estruturas Sintáticas, em 1957. Esse modelo inicial sofreu profundas modificações que culminaram com o aparecimento de *An Integrated Theory of Linguistic Descriptions*, de Katz e Postal, em 1964, e *Aspects of the Theory of Syntax*, do próprio Chomsky, em 1965. As modificações foram realmente dramáticas, uma vez que tanto a semântica quanto as considerações de natureza psicológica foram incorporadas ao processo de elaboração da teoria lingüística.

A partir desse modelo, entretanto, até o final da década de 70 e início da década de 80, a gramática gerativa sofreu fragmentações e deu origem tanto a facções teóricas como a modificações da teoria padrão que culminaram com a proposição da *Teoria da Regência e Ligação*, Chomsky (1981 e 1982) que evoluiu para a *Teoria dos Princípios e Parâmetros* (*cf.* CHOMSKY, 1986, 1988, 1989 e 1992; CHOMSKY & LASNIK, 1991).[60]

A ***Teoria Léxico-Funcional*** (BRESNAN, 1982) situa-se entre as facções que se solificaram. Essa teoria, ao lado de outras – a ***Gramática Relacional*** (PERLMUTTER & POSTAL, 1977; PERLMUTTER, 1982) e a ***Gramática Sintagmática Generalizada*** (GAZDAR, *op. cit.*) – constitui um modelo de sintaxe autônomo e alternativo. Já a ***Semântica Gerativa***, cuja origem encontra-se no trabalho de Katz & Postal (1964), teve outra sorte. Defender a centralidade da semântica dentro do "modelo padrão" parece não ter tido muito sucesso nos círculos lingüísticos, em decorrência dos contra-argumentos a essa vertente do gerativismo, levantados não só por gerativistas ortodoxos (dentre eles o próprio Chomsky e Jackendoff), que defendiam a centralidade da sintaxe, mas, principalmente, por

60 COOK (1988) apresenta uma síntese didática dos principais constructos desenvolvidos nessa teoria.

defensores da pragmática, que argumentavam contra a centralidade de ambas e, naturalmente, eram a favor da centralidade desse novo campo de investigação.

A **Teoria Léxico Funcional** (TLF) oferece uma proposta de análise lingüística adequada para fundamentar os estudos do PLN por quatro razões básicas:

- seus fundamentos formais evoluíram a partir de um modelo computacional – as *redes de transição ampliadas* (WOODS, 1970) –, e o seu procedimento formal de unificação de estruturas permite relacionar os diferentes níveis estruturais de modo algorítmico;
- sua estrutura modular prevê a integração dos vários sistemas de representação da informação lingüística: fonético-fonológico, lexical, sintático, semântico e pragmático-discursivo, possibilitando uma análise pluridimensional dos fenômenos da linguagem, imprescindível para a sua compreensão, e essencial para a implementação de SPLNs;[61]
- seus constructos permitem representar os diferentes aspectos da estrutura dos enunciados por meio de estruturas formais diferentes, refletindo, assim, as especificidades de cada nível de descrição: *estrutura de constituintes* para representar as relações superficiais de dominância e precedência dos constituintes frasais; *matrizes* de elementos hierarquizados (pares do tipo atributo-valor) para

61 Frazier (1989: 25), ao expor o conceito de "modularidade", argumenta que os sistemas de compreensão da linguagem são modulares, apresentando, portanto, vários subsistemas de processamento distintos. Cada um desses subsistemas caracteriza-se tanto por suas propriedades intrínsecas como pelas suas fontes específicas de informação. Por exemplo, o subsistema de processamento sintático pode operar sobre representações que contenham informações morfológicas, semânticas, pragmático-discursivas e, até mesmo, informações sobre o contexto situacional. No entanto, somente as informações que se sobrepuserem às suas próprias representações é que lhe são "visíveis".

representar as relações gramaticais abstratas; *esquemas funcionais* para descrever as propriedades das estruturas funcionais e para especificar as relações anafóricas e de controle; *equações funcionais* para especificar as entradas dos itens lexicais; *projeções* para estabelecer as correspondência entre as diversas estruturas sancionadas pelo modelo;

- seu tratamento diferenciado para o léxico prevê uma especificação exaustiva de informações gramaticais (forma fonética, características morfossintáticas, traços semânticos, forma semântica) para a caracterização dos itens funcionais e lexicais.

O principal objetivo da 'gramática transformacional realista', nome dado ao primeiro trabalho de Bresnan, foi, de um lado, mostrar a necessidade de aproximação entre teorias da competência – teorias que focalizam os sistemas de relações estruturais abstratas que caracterizam a língua (caracterizáveis de maneira explícita por um conjunto de princípios que especificam as frases da língua e suas estruturas fonológica, morfológica, sintática e semântica) e teorias do desempenho – teorias que focalizam os processos de compreensão e produção do discurso; de outro, lançar as bases de uma teoria lingüística capaz de 'produzir' gramáticas computacionalmente tratáveis. Como diz Bresnan, hoje, a 'arte de projetar gramáticas' pode ser considerada uma profissão. Nos laboratórios de pesquisa sobre o PLN, há uma grande demanda de gramáticas modulares, passíveis de serem representadas em alguma linguagem de programação.

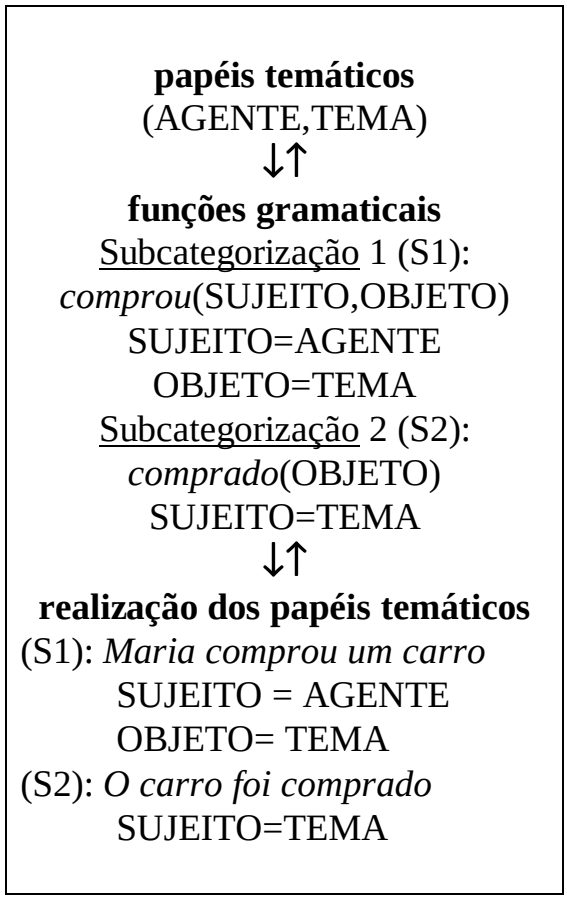
A face gramatical

Os argumentos para a proposição de uma **Gramática Léxico-Funcional** (GLF) fundamentam-se na busca de solução para as seguintes questões:

- Como projetar uma gramática capaz de representar de maneira sistemática a imensa variabilidade de expressões das línguas?
- Qual é o papel das funções gramaticais em uma teoria da gramática?

O conceito de "função gramatical" tem sido motivo de controvérsia entre os teóricos. Há lingüistas que contestam, em qualquer teoria lingüística, a inclusão de funções gramaticais como, por exemplo, a de 'sujeito' e 'objeto' por considerá-las noções, cuja especificação é dependente do processo de atribuição de papéis temáticos (FILLMORE, 1968) ou de posições configuracionais determinadas (CHOMSKY 1965 e 1986).

Para Marantz (*op. cit.*: 2), as funções gramaticais, entretanto, são mediadoras na conexão entre papéis temáticos e suas realizações. Sua maior tese foi mostrar que "as relações semânticas que se estabelecem entre morfemas ou palavras determinam as relações gramaticais de uma frase, e estas determinam a sua estrutura sintática superficial"(*op. cit.*:14), como se demonstra a seguir com o exemplo *Maria comprou um carro*:



Levando essa idéia às últimas conseqüências, Bresnan (1981, 1982 e 1988) toma, portanto, na construção da TLF, as funções gramaticais como elementos universais não definíveis.[62] Cabe, aqui, distinguir uma importante diferença entre os conceitos de função gramatical (FG) e de relação gramatical (RG). As funções gramaticais referem-se às funções sintáticas que os constituintes oracionais desempenham na frase. As relações gramaticais referem-se às associações que se estabelecem entre os constituintes oracionais e papéis temáticos. Temos, então:

62 Decisão análoga é tomada por Perlmutter & Postal, (*op. cit.*) e Perlmutter (*op. cit.*) no contexto da gramática relacional.

FRASE	***Maria*** (SN1) *comprou* ***o carro*** (SN2)
Função Gramatical	SUJEITO = SN1; OBJETO = SN2
Relação Gramatical	SN1 = AGENTE; SN2 = TEMA

A GLF prevê duas grandes classes de funções gramaticais: funções exigidas pelo predicador (funções subcategorizáveis) e funções opcionais (funções não-subcategorizáveis). A primeira classe é composta pelas seguintes funções: sujeito (SUJ), objeto (OBJ), segundo objeto (OBJ2), complemento oracional com sujeito foneticamente realizado (COMP),[63] complemento oracional com sujeito foneticamente vazio (XCOMP), complemento oblíquo (OBL_{AGENTE}, OBL_{META}, etc.) e a função de possessivo (POSS).

O traço distintivo [+ irrestrita], empregado para designar as funções gramaticais semanticamente irrestritas, subdivide a classe das funções subcategorizáveis em duas subclasses. As funções SUJ e OBJ são consideradas [+ irrestritas], as demais [- irrestritas]. As funções [+ irrestritas] possuem duas características definidoras (LEVIN, *op. cit.*: 4): (i) elas podem ser não-temáticas, como em ∅ *Parece que o vaso quebrou* e ∅ *Choveu*, e, em geral, (ii) não estão inerentemente associadas a um papel temático particular: a função SUJ pode estar associada a AGENTE, em *Pedro quebrou o vaso*, ou a TEMA, em *O vaso quebrou*, ou ainda a EXPERIENCIADOR, em *Os alunos gostaram da festa*; a função OBJ pode estar associada a EXPERIENCIADOR, em *A festa impressionou os alunos*, ou a TEMA em *Alguém deu um vaso*

63 É importante não confundir a função gramatical **COMP** com a categoria sintática **Comp** (complementador).

aos alunos. O quadro, a seguir, resume a classe das funções subcategorizáveis:

Funções Subcategorizáveis	
+ irrestrita	- irrestrita
SUBJ	OBJ2
OBJ	OBL_{θ}
	COMP
	XCOMP
	POSS

As funções gramaticais SUJ, OBJ, OBL_{θ}, em que θ designa um papel temático, e COMP são ilustradas a seguir. Dada a frase *Maria contou a Ana que Paulo colocou a faca sobre a mesa*, temos as seguintes funções gramaticais associadas aos constituintes oracionais:

SUJ: *Maria* , *Paulo*
OBJ: *a faca, a mesa*
OBL_{META}: *a Ana*
$OBL_{LOCATIVO}$: *sobre a mesa*
COMP: *que Paulo colocou a faca sobre a mesa*

A entrada lexical do verbo *preferir* ilustra a função XCOMP. A forma lexical (simplificada) é dada por:

forma lexical:[64] ‘preferir < (SUJ) (XCOMP) >’
V

↓↑ ↓↑

EXPERIENCIADOR PROPOSIÇÃO

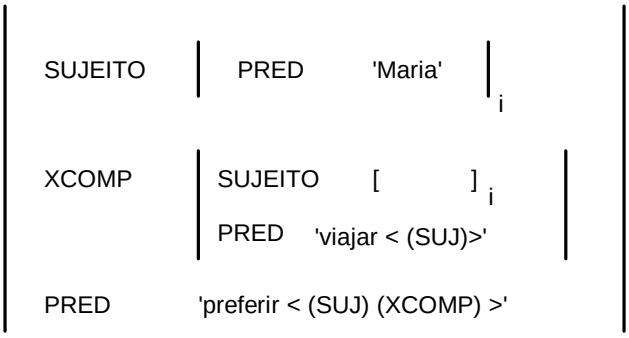
(↑ XCOMP SUJ) = (↑ SUJ)

Essa forma lexical indica que o verbo *preferir*, em frases como *Maria prefere viajar*, possui a propriedade de subcategorizar SUJ e um complemento oracional, cujo sujeito é foneticamente vazio, XCOMP. O verbo *viajar* é representado pela forma lexical ‘viajar < (SUJ) >’. O constituinte oracional *viajar* é o valor de XCOMP, que designa um “complemento de extensão oracional aberto”. Daí a representação mnemônica XCOMP, em que X representa o elemento que “falta”, para que a estrutura se torne completa, posto que o verbo viajar subcategoriza um SUJ. A regra de controle funcional, associada à entrada lexical de *preferir* (↑ XCOMP SUJ) = (↑ SUJ), fornece o processo de interpretação do sujeito de *viajar*, especificando que o valor que preenche o atributo SUJ de *viajar* deve ser preenchido pelo mesmo elemento que preenche o atributo SUJ de *preferir*, isto é, *Maria*. O processo de *unificação* (que será descrito oportunamente) produz a *estrutura funcional* da frase exemplo:[65]

---

64 Essa notação especifica que *preferir* é um verbo que subcategoriza as funções de SUJ e XCOMP, associadas aos papéis temáticos EXPERIENCIADOR e PROPOSIÇÃO respectivamente. O esquema funcional especifica que a função SUJ da oração encaixada e a função SUJ da oração matriz são preenchidas pelo mesmo elemento. Os colchetes angulares encerram os argumentos do predicador; o símbolo PRED sinaliza que essa forma possui conteúdo semântico, e as aspas simples delimitam esse conteúdo. Observe que na TLF as relações gramaticais são diretamente codificadas no léxico.

65 Os índices especificam que os atributos SUJ nas duas sub-estruturas-f possuem o mesmo valor.



A TLF atribui duas estruturas sintáticas a cada frase da língua. A primeira, chamada *estrutura de constituintes*, **estrutura-c**, equivale, aproximadamente, à estrutura de superfície, proposta pela gramática gerativo-transformacional. Essa estrutura é projetada a partir da forma fonética da frase. A segunda, chamada *estrutura funcional*, **estrutura-f**, especifica de maneira explícita as informações semanticamente interpretáveis expressas pela frase. Em outras palavras, a estrutura funcional pode ser interpretada como uma "pré-forma lógica", contendo todas as informações necessárias para a construção da forma lógica da frase: a especificação das categorias de tempo, aspecto, gênero, número, e assim por diante.

A estrutura-c e a estrutura-f são construídas a partir de *regras sintagmáticas associadas a esquemas funcionais* e de *entradas lexicais*. As regras sintagmáticas são semelhantes às regras de uma gramática gerativa, embora apresentem duas diferenças importantes: (a) elas especificam as estruturas de superfície, uma vez que a TLF não prevê dois níveis de estrutura sintagmática; e (b) cada categoria (representada no lado direito da regra sintagmática) está associada a um *esquema funcional* que especifica funções gramaticais.

No léxico, as entradas lexicais são estruturadas com as seguintes especificações básicas:[66] *forma fonética* (ou ortográfica), *categorias sintáticas* (por exemplo, substantivo, verbo, adjetivo, preposição), *funções gramaticais* (por exemplo, SUJ, OBJ, OBL), *relações gramaticais* (por exemplo, SUJ ↔ AGENTE, OBJ2 ↔ META), *atributos* (por exemplo, PRED, TEMPO, ASPECTO, NÚMERO, GÊNERO, PESSOA), *valores* (por exemplo, sg (singular), fem (feminino), 2 (segunda pessoa), nom (nominativo), pro (pronome), pas (passado)), *esquemas funcionais* (por exemplo, (↑ XCOMP SUJ) = (↑ OBJ)), *metavariáveis* (↑ e ↓), *equações funcionais* (por exemplo, (↑ TEMPO=pas)), *equações de restrição* (por exemplo, (↑ CASO) =c acus) e *formas semânticas* (por exemplo PRED 'casa', PRED 'beber <AGENTE TEMA>, 'parecer < PROPOSIÇÃO> SUJ').

As entradas lexicais dos verbos *quebrar* e *dar*, por exemplo, são representados pelas seguintes equações:

quebra- **V** (↑ PRED) = ' quebrar < AGENTE TEMA > '

↓↑ ↓↑

SUJ OBJ

da- **V** (↑ PRED) = ' dar < AGENTE TEMA META> '

↓↑ ↓↑ ↓↑

SUJ OBJ OBJ

66 Andrews (1989) apresenta as várias propostas de estruturação do léxico que surgiram a partir dos primeiros trabalhos inicialmente propostos por Chomsky, na década de 50.

Essas equações codificam, portanto, três tipos de informações distintas:

(a) o esquema de subcategorização do verbo, indicado pelas funções gramaticais (SUJ, OBJ, etc.),

(b) a estrutura de argumentos do predicador, indicado pelos papéis temáticos (AGENTE, TEMA, etc.),

(c) as relações gramaticais estabelecidas entre (a) e (b), isto é, a constituição dos pares (θx,fy) – associações estabelecidas entre um papel temático e uma função gramatical.

Os papéis temáticos funcionam como elos entre a sintaxe e a semântica. A *"Teoria da Associação entre Papéis Temáticos e as Funções Gramaticais"* (*"Linking Theory"*) especifica os modos de associação entre os papéis temáticos e as funções sintáticas (*cf.* LEVIN, *op. cit.*). Essa teoria procura propor soluções para o seguinte tipo de questão: o papel temático θx deverá estar associado a qual função sintática fy? No caso de frases passivas, a solução para essa equação seria a seguinte: o papel temático AGENTE é associado à função gramatical $\text{OBLÍQUO}_{\text{AGENTE}}$. Essa transformação é efetuada por uma regra lexical que modifica a estrutura de argumentos do predicador. Se o predicador possui a forma lexical PRED 'comprar <(↑ SUJ) (↑OBJ)> ', as regras lexicais (SUJ) → ∅ / $(\text{OBL}_{\text{AGENTE}})$, (OBJ) → (SUJ) operam sobre esssa forma tranformando-a em PRED 'comprar <$(\uparrow\text{OBL}_{\text{AGENTE}}$ (↑SUJ)'. Essa regra especifica que o argumento SUJEITO é apagado ou transformado em complemento oblíquo, desempenhando o papel temático AGENTE. É importante ressaltar que a TLF impede que qualquer regra sintática altere as correspondências entre função gramatical e papel temático. Assim, qualquer alteração desse tipo precisa

ser necessariamente feita no léxico, que passa a conter ambas as formas: a original e a transformada.

Assim, a TLF possibilita incluir no léxico informações como:

- realização fonética e gráfica dos itens lexicais;
- categorias sintáticas;
- restrições selecionais;
- formas semânticas;
- funções gramaticais;
- papéis temáticos;
- relações gramaticais;
- traços semânticos como [+/- genérico], [+\- específico], [+/- definido], [+/- humano], [+/- concreto], entre outros;
- categorias de gênero, número, pessoa, caso, voz, tempo, aspecto e modo;
- traços de reflexividade e ergatividade;

Para se ter uma visão mais concreta dos diferentes níveis de representação envolvidos, considere o verbo *comprar*, por exemplo. A ação de comprar sempre envolve um comprador, um vendedor, o objeto comprado, o preço do objeto, o dinheiro para efetuar o pagamento, a hora da compra e o local da compra, entre outros elementos. A estrutura de argumentos do verbo *comprar*, entretanto, seleciona apenas três desses elementos: o comprador, o objeto comprado e o vendedor. Os papéis temáticos selecionados são respectivamente: AGENTE, TEMA e ORIGEM. A esses papéis temáticos correspondem as funções gramaticais, que funcionam como a interface entre a sintaxe e a

semântica: SUJ, OBJ e OBL, respectivamente. Nesse nível, ocorre uma possibilidade de seleção: a forma ativa *comprar* (AGENTE = SUJ, TEMA = OBJ , ORIGEM = OBL) e a forma passiva *comprado* (AGENTE = OBL , TEMA = SUJ , ORIGEM = OBL), rebaixando, portanto, o AGENTE da transação. No próximo nível, da codificação sintática, essas funções são projetadas em posições cofiguracionais: SUJ é o primeiro nódulo dominado por F, OBJ é o primeiro nó dominado por SV e OBL é também o primeiro nó dominado por SV.[67]

O esquema abaixo resume os quatro níveis de representação.

	níveis de codificação	**expressão lingüística**: *comprar*
I	participantes do evento	(comprador,objeto comprado,vendedor)
II	estrutura temática	(AGENTE,TEMA,ORIGEM)
III	estrutura funcional	(SUJ,OBJ,OBL)
IV	estrutura sintática [68]	[SN,F] [SN,SV] [SP,VP]

Nos exemplos de regras sintagmáticas, a seguir, os esquemas funcionais estão associados às categorias sintagmáticas que aparecem do lado direito da seta $\rightarrow$. Há dois esquemas funcionais básicos:

$$(\uparrow \mathbf{FG}) = \downarrow$$

67 Não estou considerando aqui a Teoria X', (*cf.* HAEGEMAN, 1991), que projetaria as seguintes configurações: [N",Flex], [N",V'] e [N",V'].

68 Cabe observar que em línguas não-configuracionais, como é o caso do japonês e latim, por exemplo, não há projeções em termos de posições estruturais, mas sim em termos de marcas flexionais. Outra observação também importante é notar que pronomes nulo não são sintaticamente codificados.

[Lê-se: ‘‘a estrutura-f do constituinte ao qual este esquema está associado deve ser inserida como o valor da função gramatical (FG) da estrutura-f do constituinte à esquerda da regra ’’]

$$\uparrow = \downarrow$$

[Lê-se: ‘‘a estrutura-f do constituinte ao qual este esquema está associado é parte imediata da estrutura-f do constituinte à esquerda da regra’’]

F →	(SN)	SV	(SAdv)	SP*
	(↑ SUJ) = ↓	↑ = ↓	(↑ ADJUNTO)	(↑ ADJUNTO)

SV →	(CL)	V	(SV)	SP* [69]	(F')	
	(↑ OBJ) = ↓	↑=↓	(↑ OBJ) = ↓	(↑ (↓ CASO)) = ↓	(↑ COMP) =↓	
	(↓ CASO) =c ACUS					

SN →	(Det)	N	(SA)	(SP)
		↑=↓	(↑ ADJUNTO) = ↓	(↑ ADJUNTO) = ↓

SP →	P	SN
	↑=↓	(↑ OBJ) = ↓

F' →	COMP	F
		↑ = ↓

SN →	SN	F'
	↑=↓	(↑ RELATIVA) = ↓

69

$$(\uparrow (\downarrow \text{CASE})) = \downarrow \quad \Leftrightarrow \quad \begin{cases} (\uparrow \text{TO}) = \downarrow \\ (\downarrow \text{PCASE}) = \text{TO} \end{cases}$$

F' → X" F
(↑ TÓPICO) = ↓ ↑ = ↓
(↑ XCOMP* FG) = ↓

O léxico e o conjunto de regras sintagmáticas assim descritos operam em conjunto no processo de construção da ESTRUTURA DE CONSTITUINTE ANOTADA, a estrutura-c. A partir da estrutura-c constrói-se a ESTRUTURA FUNCIONAL, a estrutura-f. Em se tratando de um processo bastante trabalhoso, tomarei uma frase simples – *Pedro viu Maria* –, e, passo a passo, mostrarei como a estrutura-f é projetada a partir da estrutura-c.[70]

O léxico contém as seguintes entradas:[71]

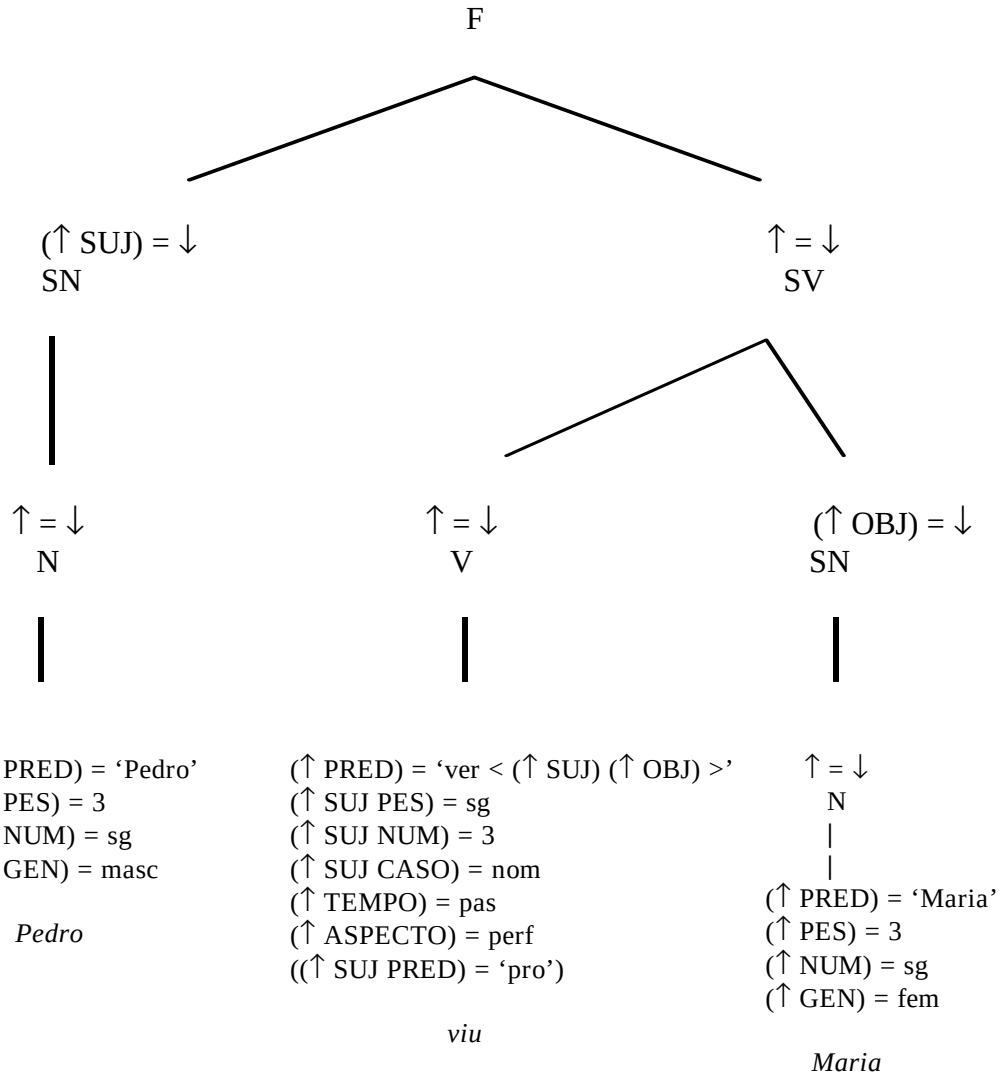
Pedro	N	(↑ PRED) = 'Pedro' (↑ PES) = 3 (↑ NUM) = sg (↑ GEN) = masc (↑ ANIM)= +
Maria	N	(↑ PRED) = 'Maria' (↑ PES) = 3 (↑ NUM) = sg (↑ GEN) = fem (↑ ANIM)= +
viu	V	(↑ PRED) = 'ver < (↑ SUJ) (↑ OBJ) >' (↑ SUJ PES) = sg (↑ SUJ NUM) = 3 (↑ SUJ CASO) = nom (↑ TEMPO) = pas (↑ ASPECTO) = perf ((↑ SUJ PRED) = 'pro')

As regras sintagmáticas e o léxico, aquelas interpretadas não como reagras de produção, mas como CONDIÇÕES SOBRE A BOA

70 Por questão de clareza, não representarei as ligações entre as funções gramaticais e os papéis temáticos.

71 A subcategorização e grade temática dos verbos fundamentam-se na proposta de BORBA (1991).

FORMAÇÃO DA ESTRUTURA SINTAGMÁTICA, ENRIQUECIDA COM OS ESQUEMAS FUNCIONAIS, sancionam a estrutura-c abaixo.

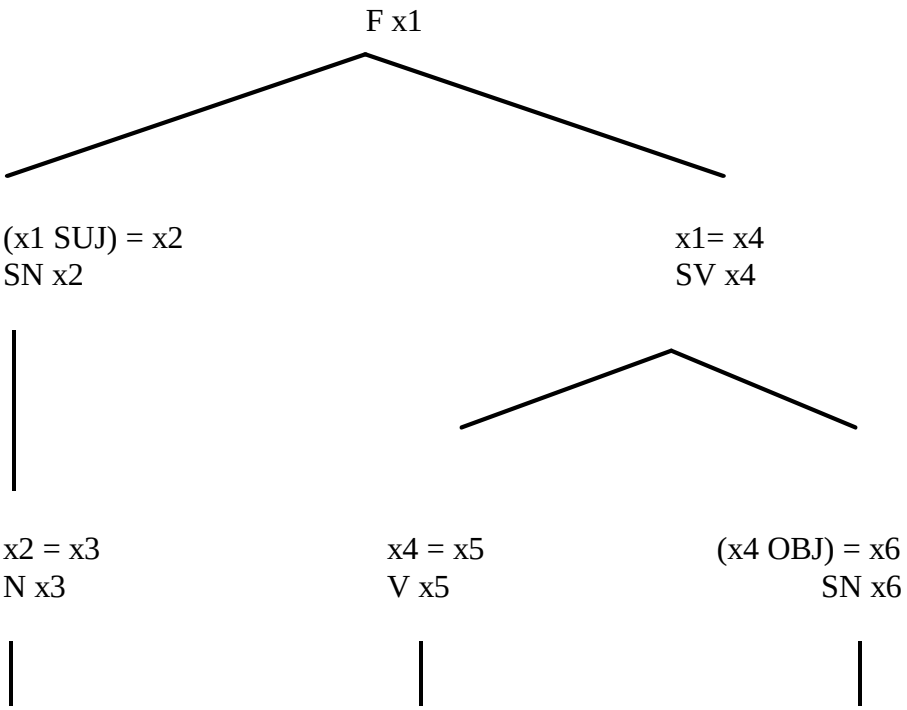


A partir dessa estrutura-c, inicia-se o processo de construção da estrutura-f. O processo consiste em substituir as metavariáveis ↑ e ↓ por váriáveis propriamente ditas. As metavariáveis servem de ponte entre as duas representações da TLF e são interpretadas como segue.

O símbolo ↓ refere-se à estrutura-f associada ao nó ao qual o esquema contendo ↓ está associado.

Já o símbolo ↑ refere-se à estrutura-f associada ao nó que imediatamente domina o nó ao qual o esquema, contendo ↑, está associado. Com essa interpretação, o processo de substituição realiza-se da seguinte maneira: para cada nó da estrutura-c, substitui-se cada uma das ocorrências da metavariável ↓ pela variável associada ao nó em que ela ocorre, e cada uma das ocorrências da metavariável ↑ pela variável do nó que imediatamente a domina. Ao final desse processo, obteremos as EQUAÇÕES FUNCIONAIS que, finalmente, servirão de base para a montagem da estrutura-f.

Cada constituinte da estrutura-c possui uma estrutura-f a ele associada. Ao se atribuirem variáveis arbitrárias (x_i), porém distintas, a cada nó da estrutura, dá-se, indiretamente, um rótulo a cada estrutura-f correspondente. Pode-se, por exemplo, fazer as seguintes atribuições:


F x1
(x1 SUJ) = x2
SN x2
x1= x4
SV x4
x2 = x3
N x3
x4 = x5
V x5
(x4 OBJ) = x6
SN x6
(x3 PRED) = ‘Pedro’
(x3 PES) = 3
(x3 NUM) = sg
(x3 GEN) = masc
Pedro
(x5 PRED) = ‘ver < (x5 SUBJ) (x5 OBJ) >’
(x5 SUJ PES) = SG
(x5 SUJ NUM) = 3
(x5 SUJ CASO) = nom
(x5 TEMPO) = pas
(x5 ASPECTO) = perf
((x5 SUJ PRED) = ‘pro’)
viu
x6 = x7
N x7
(x7 PRED) = ‘Maria’
(x7 PES) = 3
(x7 NUM) = sg
(x7 GEN) = fem
Maria

O conjunto de todas as equações funcionais, listadas abaixo, constituem a DESCRIÇÃO FUNCIONAL, a partir da qual passo a construir a estrutura-f da frase.

Descrição Funcional = conjunto de todas as equações funcionais

(x1 SUJ) = x2
x2 = x3
(x3 PRED) = ‘Pedro’
(x3 PES) = 3
(x3 NUM) = sg
(x3 GEN) = masc
x1= x4
x4 = x5
(x5 PRED) = ‘ver < (x5 SUBJ) (x5 OBJ) >’
(x5 SUJ PES) = sg
(x5 SUJ NUM) = 3
(x5 SUJ CASO) = nom
(x5 TEMPO) = pas
(x5 ASPECTO) = perf
((x5 SUJ PRED) = ‘pro’)
(x4 OBJ) = x6
x6 = x7
(x7 PRED) = ‘Maria’
(x7 PES) = 3
(x7 NUM) = sg
(x7 GEN) = fem

Cada estrutura-f possui o seguinte formato: duas colunas de elementos encerradas entre colchetes. A coluna da esquerda contém o que se denomina ATRIBUTO, a outra, o VALOR: **rótulo [ATRIBUTO VALOR]**. A colocação do rótulo da estrutura funcional é opcional. As estruturas-f parciais abaixo ilustram esses conceitos:

x3 [PRED ‘Pedro’]
x3 [PES 3]
x3 [NUM sg]
x3 [GEN masc]
x5 [PRED ‘ver < (x5 SUJ) (x5 OBJ) >’]
x5 [SUBJ PES sg]
x5 [SUJ NUM 3]
x5 [SUJ CASO nom]
x5 [SUJ PRED ‘pro’]
x5 [TEMPO pas]
x5 [ASPECTO perf]
x7 [PRED ‘Maria’]
x7 [PES 3]
x7 [NUM sg]
x7 [GEN fem]

A ordem linear dos elementos é relevante para a representação, indicando que a equação **(variável ATRIBUTO) = VALOR** é verdadeira. Em outras palavras, a equação funcional **(variável ATRIBUTO) = VALOR** resulta na estrutura funcional verdadeira **variável [ATRIBUTO VALOR]**.

Observo que os atributos são sempre símbolos simples. Já os valores podem ser símbolos simples, outras estruturas-f ou formas semânticas. Por exemplo, as estruturas-f correspondentes às equações (x5 SUJ PRED) = ‘pro’ e (x5 SUJ PES) = sg são dadas, respectivamente, por:

x5[SUJ [PRED ‘pro’]] e x5[SUJ [PES sg]]

As estruturas-f que possuem o mesmo rótulo são de fato partes de uma mesma estrutura-f. Logo, podem ser agrupadas:[72]

x3	PRED	'Pedro'		x7	PRED	'Maria'
x2	PES	3		**x6**	PES	3
	NUM	sg			NUM	sg
	GEN	masc			GEN	fem
	CASO	nom				

x5	SUJ	**x1**	PES	sg	
x4			NUM	3	
			PRED	'pro'	
	TEMPO		pas		
	ASPECTO		perf		
	PRED		'ver < (x5 SUJ) (x5 OBJ) >'		

As equações da forma xi (ATRIBUTO) = xj especificam que a estrutura-f rotulada xj deve ser inserida como o valor da estrutura-f rotulada xi. As equações x1 (SUJ) = x2 e x4 (OBJ) = x6 especificam:

(x1 SUJ) = x2

(x4 OBJ) = x6

Essas equações determinam que a estrutura-f rotulada x2 deve ser inserida como valor da estrutura-f rotulada x1 e que a estrutura-f rotulada x6 deve ser inserida como o valor da estrutura-f rotulada x4, respectivamente. Na situação (x4 OBJ) = x6, basta preenchê-la com a

[72] Por questões tipográficas, vou empregar barras verticais e não colchetes para a representação das estruturas-f mais complexas.

estrutura-f rotulada x6. Na situação (x1 SUJ) = x2, porém, o valor do atributo SUJ já se encontra parcialmente preenchido por uma estrutura-f:

SUJ	PES	sg	
	NUM	3	
	PRED	'pro'	

Isso significa que o valor final do atributo é resultante da ***unificação*** de duas estruturas-f.

A operação de ***unificação*** das expressões simbólicas, empregadas pela GLF é assim definida:

> "A unificação de duas instâncias de símbolos atômicos idênticos resulta no mesmo símbolo atômico. Duas restrições são impostas à operação: símbolos atômicos distintos não se unificam e formas semânticas (sempre representadas entre aspas simples) também não se unificam. Dadas duas estruturas x1 e x2, escolhe-se, por exemplo, x1 e, para cada atributo A de x1, procura-se uma instância de A em x2. Chamemos V o valor do atributo A de x1. Se A não ocorrer em x2, acrescenta-se o atributo A e seu respectivo valor V em x2. Caso contrário, se A ocorrer em x2, e se o valor de A em x2 for V', então a unificação de V com V' passará a ser o novo valor de A em x2. Se todas as unificações subseqüentes forem bem sucedidas, a versão modificada de x2 representará a unificação das estruturas x1 e x2, indicando que a estrutura é bem-formada. Se alguma das unificações não for bem sucedida, a estrutura, então, será mal-formada, indicando que a frase da qual a estrutura é parte é agramatical" (WESCOAT & ZAENEN, s/d: 9).

Exemplifiquemos. Dadas as estruturas-f [NUM SG] e [PES 3] a unificação dessas duas estruturas produz como resultado a estrutura:

NUM	sg
PES	3

As estruturas [NUM sg] e [NUM sg] unificam-se: [NUM sg]. Já as estruturas [NUM sg] e [NUM pl] não se unificam porque sg e pl são símbolos atômicos distintos. As estruturas [PRED 'pro'] e [PRED 'Pedro'] também não se unificam. Ambos os valores das instâncias dos atributos PRED são formas semânticas únicas.

Aplicando-se, finalmente, a operação de unificação ao nosso exemplo inicial, obtemos:[73]

x5	SUJ	PRED	'Pedro'
x4		GEN	masc
		PES	sg
		NUM	3
		CASO	nom
	OBJ		
		PRED	‘Maria’
		PES	3
		NUM	sg
		GEN	fem
	TEMPO	pas	
	ASPECTO	perf	
	PRED	‘ver < (x5 SUJ) (x5 OBJ) >’	

Essa é a estrutura-f completa da frase *Pedro viu Maria*, que contém a informação sintática necessária para a determinação da gramaticalidade, além de fornecer também a informação relevante para o componente semântico da gramática.

As estruturas funcionais são reguladas por condições de boa-formação que filtram as estruturas sancionadas pelas regras

73 Como a especificação PRED 'pro' é opcional, a estrutura não é agramatical. Essa especificação é simplesmente ignorada.

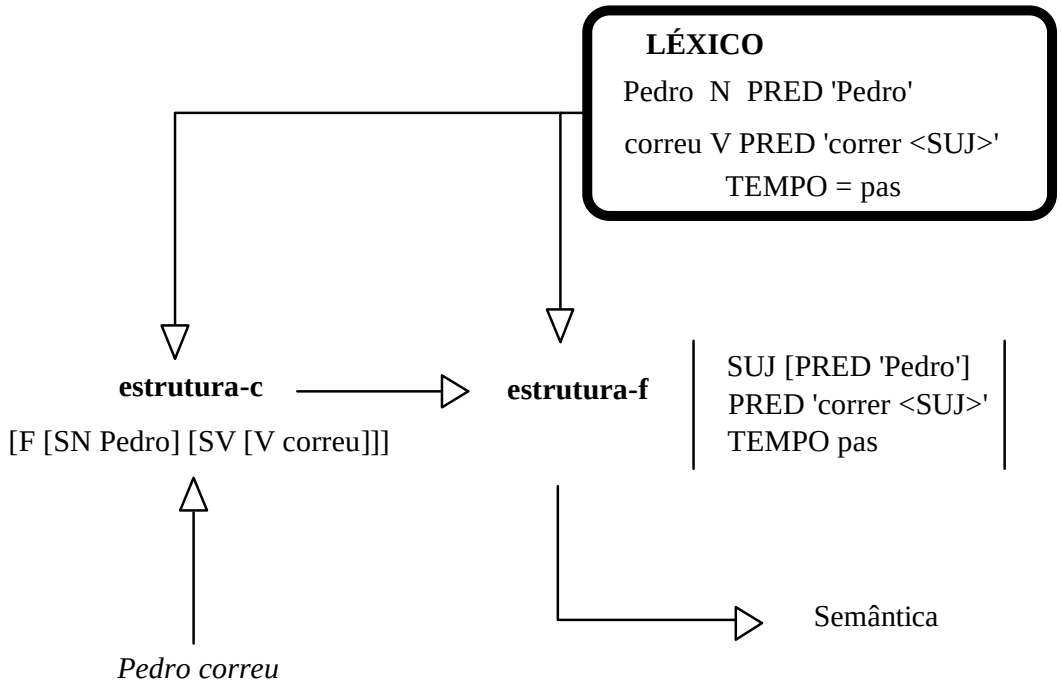
sintagmáticas que, por definição, sancionam um número muito grande de estruturas, nem todas pertencentes à língua. São elas (SELLS, *op. cit.*: 146-7):

- **Consistência** ou biunivocidade das estruturas funcionais: é uma restrição que garante o estatuto de funções atribuído às estruturas funcionais. Em outras palavras, essa condição verifica que existe uma relação de univocidade entre atributos e valores.

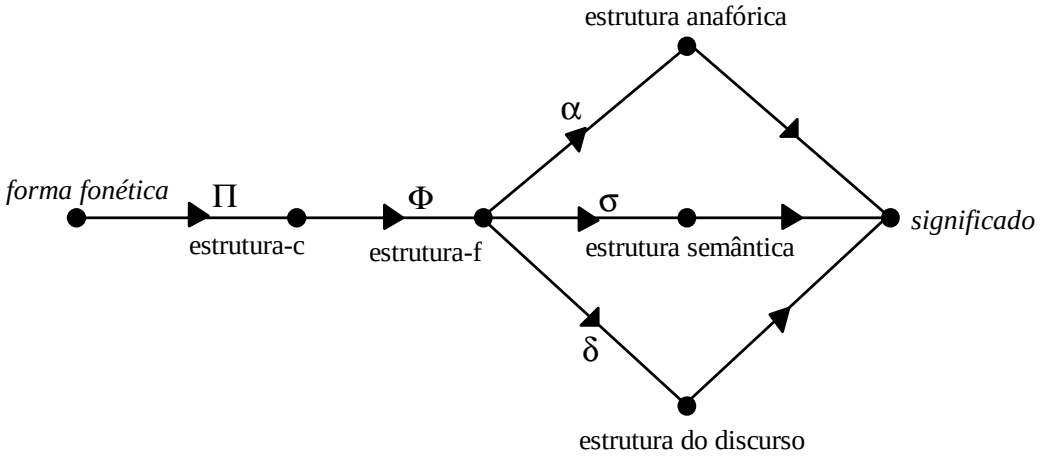
- **Completude** das estruturas funcionais: é uma restrição que garante que os esquemas de subcategorização dos predicadores sejam plenamente realizados. Em outras palavras, essa condição verifica se não há argumentos a menos.

- **Coerência** das estruturas funcionais: é uma restrição que verifica se os argumentos são de fato argumentos de um predicador. Em outras palavras, essa condição verifica se não há argumentos a mais.

O esquema a seguir sintetiza os componentes básicos e os níveis sucessivos de codificação da GLF :



Embora a TLF tenha proposto de início apenas duas estruturas, a estrutura-c e a estrutura-f, sua estrutura modular prevê a possibilidade de se construir um sistema de projeções capaz de correlacionar um complexo de informações lingüísticas: estruturas semânticas, dependências anafóricas e funções discursivas, por exemplo (DALRYMPLE, *et al.*, 1992). O esquema de Kaplan (1989: 12), a seguir, reflete a proposta de correlacionar a forma externa de um enunciado e a representação interna de seu significado:



A projeção Π, que estabelece a correspondência entre a forma fonética do enunciado e a estrutura-c, e a projeção Φ, que constrói a estrutura-f a partir da estrutura anterior, foram amplamente discutidas até este ponto.

A projeção σ nos transporta para o domínio semântico. Essa projeção possibilita acoplar um componente semântico à teoria e, a partir dele, construir a correspondência entre a estrutura-f e a estrutura semântica, que inclui a representação da forma lógica. A projeção α, também a partir da estrutura-f, possibilita a especificação de uma estrutura que registra as relações anafóricas: duas estruturas-f co-referenciais, por exemplo, seriam projetadas em um único elemento da estrutura anafórica. Por fim, a projeção δ nos leva à dimensão do discurso, que, entre outras coisas, possibilita acoplar um modelo de estruturas do discurso à teoria.

A modularidade prevista pela teoria, entretanto, precisa ser mais bem explorada, uma vez que a atenção maior tem sido dada aos aspectos sintáticos. Só muito recentemente é que tem havido tentativas de se desenvolver um modelo semântico. Neste capítulo, procuro esboçar uma possibilidade de acoplar a essa teoria o modelo semântico proposto por Jackendoff (1990) e o modelo das estruturas do discurso proposto por Grosz & Sidner (*op. cit.*). Embora não pretenda, aqui, articular esse modelo complexo, faço algumas considerações sobre uma possibilidade que me parece fecunda no contexto do PLN.

A face semântica

As teorias da significação lingüística podem ser classificadas em três grandes correntes que se complementam (*cf.* LADUSAW, 1989; CHIERCHIA & McCONNELL-GINET, 1990):

- as teorias representacionais (*cf.* JACKENDOFF, 1990), enfatizando a linguagem enquanto representação, focalizam as conexões entre a linguagem e os constructos mentais. Nesse sentido, essas teorias procuram caracterizar o conhecimento semântico que os falantes possuem de sua língua.
- as teorias referenciais (*cf.* DOWTY *et al.*, 1981; BARWISE & PERRY, 1983 e 1990; COOPER *et al.*, 1990), privilegiando o aspecto informacional da linguagem, focalizam as conexões entre a linguagem e o mundo, seja ele objetivo ou subjetivo. Nesse sentido, essas teorias procuram caracterizar as relações entre as expressões lingüísticas e as situações, sejam elas fenômenos físicos e concretos ou mentais e abstratos. Este tipo de semântica enfatiza o estudo da porção do significado que se caracteriza em termos de condições de verdade.
- as teorias sócio-pragmáticas (*cf.* AUSTIN, 1962 e 1990; GRICE, *op. cit.*; SEARLE, 1990a e 1990b; LEVINSON, *op. cit.*; LEECH, *op. cit.*; HORN, *op. cit.*), dando destaque ao aspecto da linguagem enquanto uma forma de agir sobre o mundo, focalizam o processo de comunicação verbal. Nesse sentido, essas teorias

> procuram caracterizar os atos de fala e a intenção subjacente ao processo comunicativo.

Lyons (1981) explicita a complementaridade:

> "A semântica de condição de verdade pode ser considerada complemento da pragmática. Podemos asseverar, negar ou conhecer uma proposição. Podemos também duvidar de proposições. Podemos ainda acreditar em proposições. Uma proposição pode até mesmo ser expressa por paráfrases distintas. É o seu valor-verdade, entretanto, identificado com um ou com outro dos dois valores - verdadeiro ou falso - é que é asseverado, negado ou conhecido."

Pouco se sabe sobre a natureza do significado ou como ele deve ser representado. Há, entretanto, propostas que procuram caracterizá-lo. Uma das primeiras propostas foi representá-lo em termos de primitivos ou traços semânticos (KATZ, *op. cit.*; JACKENDOFF, 1972). O significado do item lexical *cadeira*, por exemplo, seria o conjunto de primitivos:

[*cadeira*] = {OBJETO, CONCRETO, INANIMADO, ARTEFATO, MOBÍLIA, TRANSPORTÁVEL, COM PERNAS, COM ENCOSTO, COM ASSENTO, ASSENTO INDIVIDUAL}.

Esse enfoque não está livre de problemas, porque é extremamente difícil, talvez até impossível, delimitar uma coleção universal de primitivos semânticos, a partir da qual os significados de todos os itens lexicais (de todas as línguas) possam ser especificados (KEMPSON, 1977). No exemplo acima, há primitivos que não são exclusivos do significado de 'cadeira', e outros nem mesmo se aplicam a determinados tipos de 'cadeiras'. Outro problema é como determinar o

significado das frases a partir desses primitivos. Em outras palavras, como integrar sintaxe e semântica.

Reinterpretando a idéia de primitivos semânticos e procurando resolver essas dificuldades, Jackendoff (1990) constrói um modelo semântico em que o problema da coleção universal de primitivos semânticos é contornado, pois postula um conjunto de *categorias ontológicas*: COISA, EVENTO, ESTADO, AÇÃO, LUGAR, TRAJETÓRIA, PROPRIEDADE e QUANTIDADE. Essas "partes conceituais do discurso" constituem as unidades essenciais da *estrutura conceitual* – nível de representação mental em que são codificadas as representações do mundo.

Postulando a existência de quatro domínios – "mundo real", "mundo projetado" ou "mundo da experiência", "estrutura conceitual" e "expressões lingüísticas" –, Jackendoff (1983: 31) propõe uma sistematização que tem por objetivo (i) caracterizar o tipo de objeto pertencente a cada um desses domínios e (ii) equacionar as inter-relações entre esses domínios. O quadro, a seguir, apresenta uma síntese de sua proposta:

mundo real	**mundo projetado**	**estrutura conceitual**	**expressão lingüística**
(ENTIDADE)	(REFERENTE)	(SENTIDO)	(FORMA)
[objetos físicos]	[objetos percebidos]	[objetos mentais]	[objetos lingüísticos]
cor	# cor #	COR	*cor*
luz	# luz #	LUZ	*luz*
calor	# calor #	CALOR	*calor*

Esse quadro possibilita a seguinte leitura: cor, luz e calor são diferentes tipos de radiações eletromagnéticas (objetos físicos), que experienciamos enquanto #cor#, #luz# e #calor# (objetos percebidos), repectivamente. Cada um deles é, portanto, parte do mundo projetado, resultante do processo de "filtragem" da realidade. Já os objetos COR, LUZ e CALOR são estruturas conceituais (objetos mentais), construídas a partir da percepção dos objetos físicos e responsáveis pela criação do mundo projetado. Por fim, *cor*, *luz* e *calor*, são, neste caso, representações ortográficas de seqüências de sons com a propriedade de veicular informação (objetos lingüísticos).

O problema da relação entre a sintaxe e semântica é, então, formalizado com a proposição de *regras de projeção* que estabelecem as correspondências entre as duas estruturas: conceitual e sintática. Observe o exemplo, adaptado de Jackendoff (*op. cit.*: 45):

FRASE: *Pedro correu*
ENTRADAS LEXICAIS:[74] **Pedro** N $[_{OBJETO}$ PEDRO$]$ **correr** V ____ < SP_j > $[_{EVENTO}$ IR ($[_{OBJETO}$ $]_i$,$[_{TRAJETÓRIA}$ $]_j$)$]$
Estrutura sintática:

[74] Os índices i e j correspondem aos argumentos externo (SUJ) e interno (LOC), respectivamente.

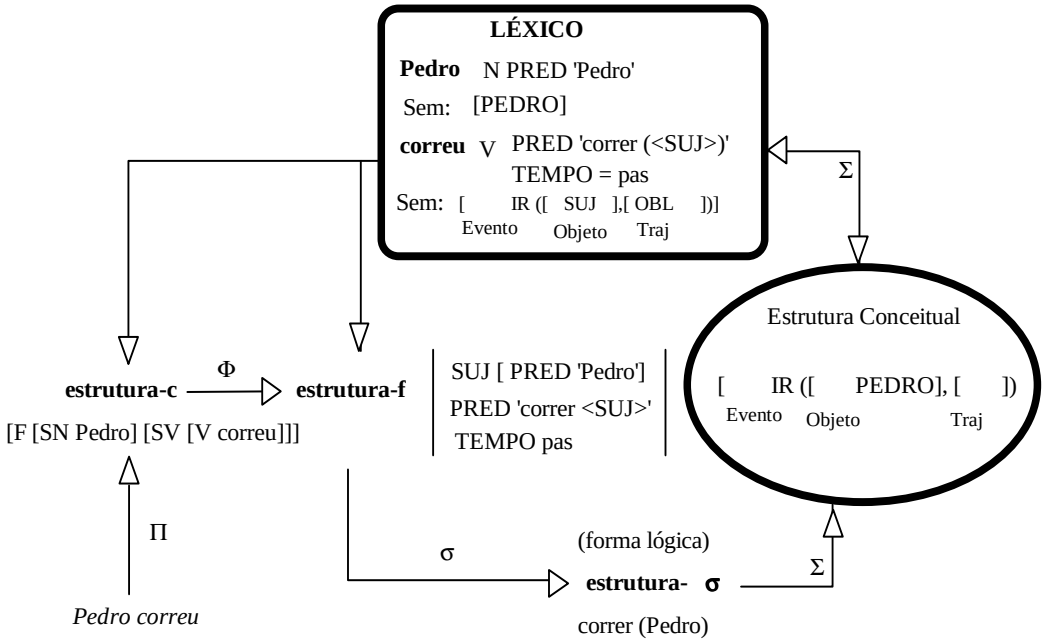
$[_F\ [_{SN}$ Pedro$]\ [_{SV}$ correu $]\ ]$
Estrutura semântica: $[_{EVENTO}$ IR $([_{OBJETO}$ PEDRO$],[_{TRAJETÓRIA}\ \varnothing\ ]_j)]$

Observe que Jackendoff decompõe o significado do verbo *correr* em termos dos conceitos EVENTO, IR e OBJETO e propõe uma correspondência entre as estruturas sintática e conceitual:

Estrutura sintática	**↔**	**Estrutura conceitual**
Frase	↔	EVENTO
SUJ	↔	OBJETO
∅	↔	TRAJETÓRIA
V	↔	$[_{EVENTO}$ IR ([OBJETO] , [TRAJETÓRIA])]

A proposta de Jackendoff e a da TLF podem se completar, uma vez que a primeira fornece o elo de ligação entre as entidades lingüísticas e o nível conceitual (as regras de correspondência Σ) e a segunda estabelece a ponte entre as entidades lingüísticas e sua forma fonética (as projeções Π, Φ e σ). A forma semântica da TLF, que é simplesmente uma especificação provisória, à espera de um módulo semântico que a desenvolva, pode ser enriquecida com as estruturas semânticas projetadas por Jackendoff.

Proponho o esquema abaixo para esboçar uma possível integração entre as propostas:



Esse esquema sugere que, dada a frase *Pedro correu*, a projeção Π, por meio das regras sintagmáticas e do léxico, projeta a estrutura-c; a projeção Φ, por meio dos esquemas e equações funcionais e do processo de unificação, produz a estrutura-f; a projeção σ, por meio das informações contidas na forma semântica, projeta a estrutura-σ, que corresponde à forma lógica "correr(Pedro)". Esta última é, então, interpretada pelas regras de projeção Σ que fazem a correspondência entre a forma lógica e a estrutura conceitual.

A face pragmático-discursiva

Do ponto de vista gramatical, as línguas humanas são analisadas como sistemas, compreendendo diversos níveis de abstração: o fonológico, o morfológico, o sintático e o semântico. Unidades de representação diversas – traços, fonemas, morfemas, palavras, sintagmas, frases, conceitos e regras – têm sido propostas para capturar

padrões e generalizações em cada um dos níveis. Do ponto de vista funcional, as línguas podem ser concebidas como complexos sistemas de comunicação, capazes de preencher uma grande variedade de funções comunicativas como, por exemplo, cumprimentar, nomear, referir, solicitar, perguntar, asseverar, descrever, prometer, negar, entre outras.

Assim, além de informações tipográficas, ortográficas, morfológicas, sintáticas e semânticas, necessárias para a construção do significado descontextualizado, observa-se que o sistema precisa também de informações para poder construir a interpretação das "frases em contexto", isto é, a interpretação dos enunciados, elementos necessariamente dependentes dos contextos pragmático-discursivo e situacional. Em outras palavras, o SPLN precisa, de alguma forma, também considerar a dimensão pragmático-discursiva, que exige, além da representação do discurso e de sua manifestação em textos, a representação de seus participantes, com suas "visões de mundo", e a especificação de conceitos que possibilitem, pelo menos, a representação de parcelas do mundo, isto é, de parcelas de "conhecimento de mundo(s)", conhecimento que se refere não só a situações particulares, mas também ao conhecimento de convenções sociais gerais e leis físicas, entre outros. Em suma, a interpretação de um diálogo, por exemplo, precisa levar em conta os enunciados que o precedem, a localização espaço-temporal em que o diálogo ocorre e os seus participantes, com suas crenças, desejos, intenções, etc..

Como apontam os filósofos, o uso da linguagem pode ser concebido como ação. As pessoas fazem uso da língua para alterar o mundo que as rodeia. Sendo assim, a compreensão dos enunciados, produzidos pelos falantes, depende também do reconhecimento das

intenções a eles subjacentes. Um falante pode, por exemplo, produzir uma seqüência de enunciados com a intenção de afetar as crenças de seu interlocutor, alterando, portanto, as próprias intenções do interlocutor, bem como outros aspectos de seu estado mental. Logo, os atos de fala – pedidos, promessas, ameaças, por exemplo – e as intenções subjacentes a eles constituem a base do processo de comunicação.

O modelo tradicional de comunicação (*cf.* DUBOIS *et al*. *op. cit.*: 129-33), proposto pela Teoria da Informação, postula que a comunicação lingüística é possível, porque o significado das expressões lingüísticas, empregadas na construção da mensagem, é tacitamente partilhado pelos participantes do evento comunicativo. Como esse modelo identifica o conteúdo da mensagem com a intenção comunicativa do emissor, determinar tal intenção é o mesmo que determinar o significado das expressões lingüísticas por ele empregadas no processo de construção da mensagem. Esse modelo de comunicação que concebe a línguagem como uma espécie de "fio condutor de idéias", entretanto, apresenta inadequações ao desconsiderar questões que são fundamentais para o equacionamento do processo de comunicação.

Quando no início deste capítulo, eu decido asseverar por escrito que o atleta do exemplo está aliviado, esse "evento comunicativo" envolve, pelo menos, quatro tipos de entidades distintas (KRONFELD, *op. cit.*: 17), em quatro níveis distintos: (nível morfossintático) a frase declarativa *Joaquim Cruz está aliviado*; (nível semântico) que expressa um conteúdo proposicional (a proposição "ALIVIADO(JOAQUIM CRUZ)"); (nível cognitivo) externa uma atitude proposicional (a minha crença [JOAQUIM CRUZ ESTÁ ALIVIADO]); e (nível pragmático) realiza um ato de fala (asseverar que Joaquim Cruz

está aliviado). Assim, a frase J*oaquim Cruz está aliviado* é a forma lingüística que selecionei para produzir um ato de fala, expressar uma proposição e veicular uma crença sobre um estado de mundo.

Observe que cada entidade envolvida nesse evento é categorialmente diferente das demais: uma seqüência de letras e sinais (a enunciação da frase declarativa escrita), uma abstração teórica (a proposição), um estado mental (minha crença) e um ato de fala (a asserção). Cada uma dessas entidades, por sua vez, pode ser generalizada para incluir outros tipos. Além de frases declarativas, por exemplo, incluem-se também os outros tipos de frases (interrogativas, negativas, imperativas); além de crenças, incluem-se também outras atitudes proposicionais (desejos, intenções, esperanças); além de asserções, incluem-se outros atos de fala (questionar, solicitar, ordenar, advertir, cumprimentar, entre outros). A partir dessas generalizações podemos concluir que todo evento comunicativo pode ser analisado em quatro dimensões: lingüística, lógica, mental e pramática.

O problema de relacionar a expressão lingüística *Joaquim Cruz* ao indivíduo único Joaquim Cruz pode ser formalizado como um ato de fala particular, envolvendo também essas quatro dimensões (*cf*. KRONFELD, *op. cit*.):

DIMENSÕES	**TIPOS**	**OBJETOS**
lingüística	expressão referencial	*Joaquim Cruz*
lógica	argumento da proposição	ALIVIADO(JOAQUIM CRUZ)
mental	representação mental	# Joaquim Cruz #
pragmática	ato de referir	*Joaquim Cruz* ↔ # Joaquim Cruz #

Diante desses fatos, Akmajian *et al*. (*op. cit.*: 398-9) argumentam que, além de partilharem uma língua comum, os participantes de um evento comunicativo partilham também de um *sistema de crenças* e *de inferências* que funciona como *estratégias comunicativas*. Defendem, em oposição ao modelo tradicional (centrado na mensagem) um "modelo inferencial de comunicação" que (i) incorpora a noção de intenções comunicativas; (ii) não reduz a interpretação das intenções comunicativas ao significado das expressões contidas nos enunciados; e (iii) fornece condições para a interpretação dos diferentes modos de comunicação: literal, não-literal, direto e indireto.

Nesse modelo, concebe-se a comunicação lingüística como um processo de reconhecimento de intenções comunicativas: o emissor e o receptor partilham um sistema de estratégias de inferências (estratégias pragmáticas) que permite ao receptor estabelecer relações entre o significado das expressões lingüísticas e as intenções comunicativas do emissor que as produziu. Em outras palavras, esse modelo estabelece que a mensagem e o significado das expressões lingüísticas estão relacionados por meio de um seqüência de inferências.

A comunicação lingüística passa a ser concebida, então, como um caso particular de **resolução de problemas**, e não como uma simples troca de mensagens – o emissor enfrenta o problema da construção da mensagem. Dado o contexto situacional em que o evento comunicativo ocorre, o emissor deve fazer com que o receptor reconheça suas intenções comunicativas, ou parte delas. Nessa tarefa, ele deve, então, construir um **plano** para alcançar seus objetivos (fazer com que o receptor da mensagem, por exemplo, fique a par de um fato), escolhendo

as expressões lingüísticas que julgar apropriadas para que sua intenção comunicativa, ou parte dela, seja reconhecida pelo receptor. O receptor, por sua vez, enfrenta o problema inverso: reconhecer a intenção comunicativa e o próprio plano do emissor, ou parte dele, em função das expressões por ele escolhidas e do contexto situacional em que o evento comunicativo ocorre (*cf.* ALLEN & PERRAULT *op. cit.*; APPELT, *op. cit.*; COHEN & PERRAULT, *op. cit.*). Assim, o objetivo da **análise** (recepção do texto) é relacionar as expressões lingüísticas às suas interpretações pragmáticas possíveis, e da **síntese** (produção do texto), selecionar, entre as expressões lingüística possíveis, aquela que melhor atinge o resultado pragmático pretendido.

A resolução desses problemas, por sua vez, depende de um conhecimento prévio, adquirido durante o processo de aprendizado (desenvolvimento) do próprio sistema lingüístico – aprende-se uma língua e aprende-se a se comunicar nessa língua. Esse conhecimento inclui expectativas partilhadas pelos indivíduos de uma comunidade lingüística:

Quadro de Expectativas Compartilhadas

Pressupõe-se que:

- o receptor seja capaz de determinar o significado e a referência das expressões lingüísticas;
- o emissor deva possuir uma intenção comunicativa identificável, a menos que haja alguma evidência ao contrário;
- o emissor e o receptor estejam empregando as expressões no seu sentido literal, a menos que haja alguma evidência ao contrário;
- o emissor observe os princípios conversacionais – Relevância, Sinceridade, Veracidade, Quantidade, Qualidade.

Além dessas expectativas, Akmajian *et al.* sugerem um sistema de estratégias de inferências, também compartilhado pelos participantes de um evento comunicativo, que norteia todo o processo de comunicação. Esse sistema é composto pelas seguintes estratégias:

Sistema de Estratégias Compartilhadas

Estratégia Direta: o receptor [1° passo] reconhece as formas lingüísticas; [2° passo] reconhece qual é o significado pretendido pelo emissor; [3° passo] determina os referentes pretendidos pelo emissor; [4° passo, e final, se a comunicação for simplesmente direta] reconhece o que o emissor está comunicando algo de modo direto;
Estratégia Literal: o receptor [5° passo] reconhece que o contexto exige que o emissor **seja literal**; [6° passo, e final, se a comunicação for direta e literal] reconhece o que o emissor está comunicando algo de modo direto e literal;
Estratégia Não-Literal: o receptor [5° ' passo] reconhece que o contexto exige que o emissor **não seja literal**; [6° ' passo, e final, se a comunicação for direta e não-literal] reconhece que o emissor está comunicando algo de modo direto, porém, não está sendo literal.
Estratégia Indireta: o receptor [7°] reconhece que o contexto exige que o emissor seja **indireto**; [8°] reconhece que o emissor está também comunicando algo de modo indireto.

Além de informações de natureza pragmática, a compreensão global de um texto pressupõe a compreensão de relações que se estabelecem entre suas partes. Como já se sabe, essas partes não são exclusivamente constituídas de frases isoladas. No texto, as frases interligam-se umas às outras para formar unidades textuais maiores, e estas unidades, por sua vez, podem também se combinarem para compor outras unidades ainda maiores. Determinar, portanto, os limites dessas unidades é tarefa essencial, uma vez que um texto pode asseverar relações semânticas específicas entre suas unidades constituintes: o conteúdo de um determinado segmento de texto pode asseverar, por exemplo, uma série de conseqüências que são decorrentes do segmento de texto adjacente. Além disso, a estrutura do texto afeta a interpretação de unidades da dimensão frasal, definindo um contexto semântico dentro do qual elementos dêiticos, anafóricos e descrições definidas, entre outros, encontram sua interpretação.

Não é fato novo afirmar quão difícil é a identificação de "descrições", "explicações", "histórias", "planos", entre outras unidades estruturais de um discurso. Tais unidades, que estavam presentes no momento da interação verbal, durante a fase de análise, parecem ter simplesmente desaparecido.

O problema da segmentação em termos de unidades discursivas é ilustrado por meio do texto, adianter, que registra a interação verbal entre cinco pessoas.

Nesse texto, os dois "protagonistas" do diálogo, A e B, estão jogando *Civilization*, um jogo de estratégia bélica. No segmento transcrito e adaptado, eles estão planejando atacar a Europa. C e D, seus adversários, estão estudando sua própria jogada enquanto aguardam a vez. Num certo momento, E, que não está participando do jogo, aproxima-se do jogador C e pergunta por alguém. Apesar da aparente "desordem" , nota-se que os participantes do evento comunicativo são capazes de acompanhar o assunto e de interagir verbalmente uns com os outros, produzindo enunciados apropriados à situação comunicativa, demonstrando uma compreensão de todas as expressões que apresentam um significado subespecificado:

A: Estamos na África, certo? Então, vamos primeiro atacar Portugal. Para falar a verdade, eu gosto de Portugal. Quando estive lá o ano passado, tomei muito vinho. E depois, a Espanha. Já te falei do restaurante que fomos em Madri?

B: Já. Acho que sim. Era muito melhor do que aquele que nós fomos em Barcelona antes de pegar o avião… Mas eu acho que não. Vamos atacar a Itália depois. Depois…

C: Dá para vocês falarem mais baixo?

D: Vocês estão atrapalhando nosso raciocínio.

A: Tá legal. Desde que a gente não precise atacar Roma.

B: Depois a Alemanha.

A: Quando nós vamos enfrentar a França? A gente não pode esquecer.

B: Depois de atacar…

E: Com licença. Vocês viram o Pedro?

C: Eu não vi. Depois quero falar com você. Por falar nisso, vocês ainda estão falando muito alto.

A: E daí?

B: Que cerveja tomamos o ano passado! E a Itália?

A: Depois de atacar a Alemanha. Você vai tirar férias este ano? Ou vai ficar flauteando no trabalho como sempre?

B: Ainda não decidi, e você?

A: Acho que vou para os Estados Unidos de novo. O próximo alvo será a Inglaterra, certo?

Por intuição, o segmento do diálogo apresenta partes em que A e B estão jogando – planejando os ataques – e partes em que A e B fazem comentários sobre locais em que estiveram em suas viagens, "jogam conversa fora" ou respondem aos adversários. Num determinado momento, A e B interrompem o planejamento estratégico do jogo para ouvir a intervenção de E.

Observe, abaixo, a disposição gráfica dos enunciados, de acordo com a seguinte hierarquia: os enunciados escritos mais à esquerda da página referem-se ao "planejamento de ataques", que constitui o assunto central; os enunciados não-centrais – comentários ou interrupções, por exemplo – estão graficamente representados mais à direita em relação aos enunciados centrais (numerados de 1 a 12)

Observe também que é possível especificar uma hierarquia entre os enunciados não-centrais. Finalmente, observe que A e B sempre

retomam o seu "planejamento de ataques" – seu foco de atenção –, depois das interrupções e comentários.

A: 1.Estamos na África, certo?
2.Então, vamos primeiro atacar Portugal.
Para falar a verdade, eu gosto de Portugal.
Quando estive lá o ano passado, tomei muito vinho.
3.E depois, a Espanha.
Já te falei do restaurante que fomos em Madri?

B: Já. Acho que sim.
Era muito melhor do que aquele que nós fomos em Barcelona antes de pegar o avião…
Mas, eu acho que não.
4.Vamos atacar a Itália em seguida.
5.Depois…

C: Dá para vocês falarem mais baixo?

D: Vocês estão atrapalhando nosso raciocínio.

A: Tá legal.
6.Desde que a gente não precise atacar Roma.

B: 7.Depois a Alemanha.

A: 8.Quando nós vamos enfrentar a França?
A gente não pode esquecer.

B: 9.Depois de atacar…
E: Com licença.
Vocês viram o Pedro?

C: Eu não vi.
Depois quero falar com você.
Por falar nisso, vocês ainda estão falando muito alto.

A: E daí?

B: Que cerveja tomamos o ano passado!
10.E a Itália?

A: 11.Depois de atacar a Alemanha.

Você vai tirar férias em este ano?
Ou vai ficar flauteando no trabalho como sempre?

B: Ainda não decidi, e você?

A: Acho que vou para os Estados Unidos de novo.
12.O próximo alvo será a Inglaterra, certo?

Tomando como ponto de partida fenômenos como esses, Grosz & Sidner (*op. cit.*) propõem um modelo formal que possibilita o tratamento computacional da estrutura do discurso que poderíamos denominar "discurso voluntário e consciente", uma vez que se pressupõe que ato comunicativo é um ato intencional.[75] Postula-se que a estrutura do discurso é resultante da interação de três estruturas: **estrutura intencional** (EI), **estrutura lingüística** (EL) e **estado de atenção** (EA).[76] A decomposição da estrutura do discurso em três componentes inter-relacionados visa a explicar a coerência global e local do discurso, as interrupções, o uso de certos tipos de expressões referenciais, a segmentação do discurso, as relações anafóricas interfrasais, entre outros fenômenos.

A EI, hierárquica por definição, descreve as relações entre a **intenção do discurso** (ID) como um todo e a **intenção de cada um de**

---

75 Há que se ressaltar que Grosz & Sidner (*op. cit.*: 176) se propõem a analisar discursos "orientados" para o preenchimento de algum tipo de objetivo: seja ele físico, mental ou lingüístico. O termo genérico "intenções" é, aqui, empregado para designar esses "objetivos" subjacentes a esses discursos. Para não enveredarmos em discussões filosóficas sobre *intenção* e *intencionalidade*, que me parecem pouco pertinentes para a simulação computacional dos tipos de discurso aqui considerados, observo, que em um quadro teórico bastante diverso, autores como Greimas & Coutés (1979: 237-38) criticam o uso generalizado do termo *intenção* – para eles, o conceito de *intencionalidade*, embora não se identifique com "motivação" e nem com "finalidade", subsume ambos. Já a noção de *intenção*, argumentam, nos leva a encarar a comunicação como um "ato voluntário" e, ao mesmo tempo, "consciente".

76 Esse modelo evoluiu a partir de Grosz (1977 e 1986), Grosz *et. al.* (1983) e Sidner (1983).

seus segmentos (ISD). Note que esta estrutura é a projeção dos atos de fala na dimensão do discurso. São exemplos de ID ou ISD: “Querer que o(s) interlocutor(es) [i] execute(m) uma determinada tarefa física, [ii] acredite(m) em um fato específico, [iii] conheça(m) uma determinada propriedade de um objeto, [iv] queira(m) identificar um objeto”. Observe:

[i] querer que Pedro abra a porta;

[ii] querer que Pedro acredite que a porta está trancada;

[iii] querer que os alunos saibam que a aula será importante;

[iv] querer que os professores queiram pesquisar mais.

O modelo prevê que essas intenções podem estar ligadas entre si por meio de relações de dominância (situação em que o preenchimento de um determinado objetivo comunicativo global depende do preenchimento de objetivos comunicativos parciais) e precedência (situação em que o preenchimento de um determinado objetivo comunicativo depende do preenchimento de um objetivo comunicativo anterior). Essas relações se assemelham às relações estruturais elementares de dominância (não-imediata) e precedência linear que se estabelecem entre os constituintes oracionais. Qualquer uma dessas intenções podem constituir-se em uma intenção global ou parcial

A EL é resultante da estrutura formada pelos enunciados que compõem um discurso. A motivação para a proposição dessa estrutura na dimensão discursiva é a estruturação observada na dimensão frasal, em que as categorias sintáticas agrupam-se para formar

constituintes oracionais, que, por sua vez, desempenham funções específicas no interior da frase. De modo semelhante, os enunciados que compõem um discurso estruturam-se para formar segmentos de discurso que, por sua vez, desempenham funções específicas em relação ao discurso como um todo. Há situações em que enunciados consecutivos não se encontram no interior de um mesmo segmento, e situações em que enunciados linearmente distantes encontram-se precisamente no mesmo segmento.

A EL é, então, formada por segmentos de discurso e de relações de encaixe que se estabelecem entre esses segmentos. Tais relações são interpretadas como reflexo superficial das relações subjacentes que se estabelecem entre os elementos da estrutura intencional. É importante observar que não se trata de uma estrutura estritamente composicional: um determinado segmento pode incluir uma combinação de subsegmentos e de enunciados contidos em si mesmo e não, nos seus subsegmentos. Observe que há uma interação particular entre a estrutura de um segmento e os enunciados que compõem o discurso: as expressões lingüísticas podem ser empregadas para veicular informação sobre a estrutura do discurso e esta restringe a interpretação das expressões. De fato, as expressões lingüísticas estão entre os principais indicadores que assinalam os limites de um segmento de discurso. O uso explícito de expressões como *em primeiro lugar*, *por fim*, *assim*, *para falar a verdade*, entre outras, e as mudanças de tempo e aspecto são dispositivos lingüísticos que, sem dúvida, podem funcionar como marcadores discursivos, isto é, como sinalizadores da estrutura do discurso que se desdobra.

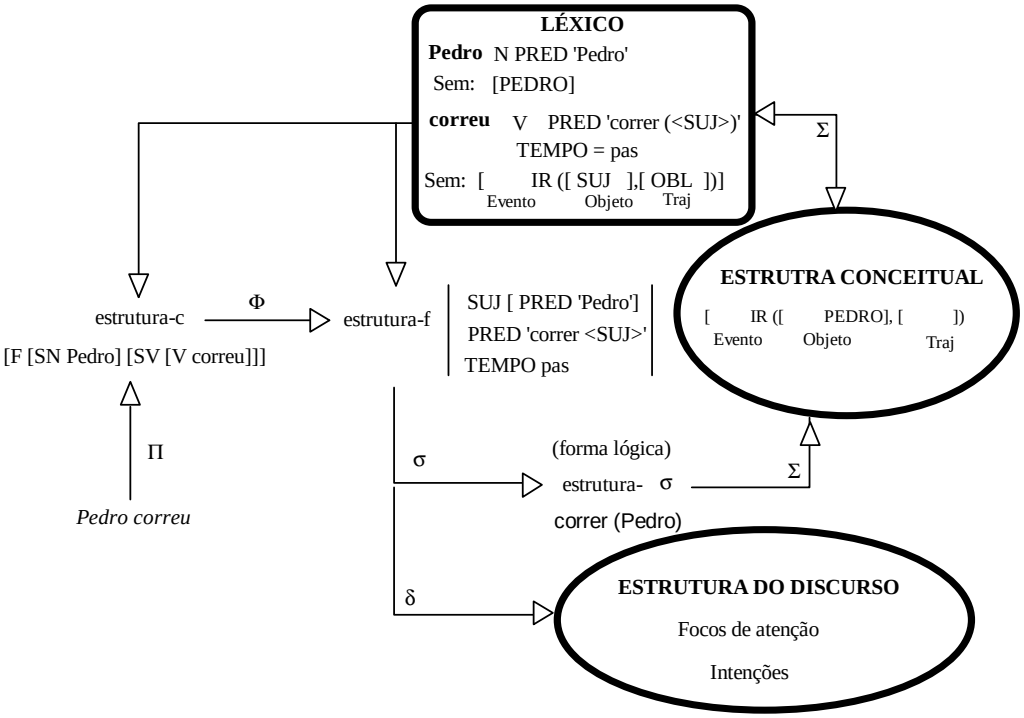
O EA representa dois tipos de *foco de atenção* dos participantes do evento comunicativo, à medida que o discurso se desdobra: *foco imediato*, que opera na dimensão do enunciado, e *foco global*, que opera na dimensão dos segmentos do discurso. Grosz e Sidner enfatizam que o estado de atenção é uma propriedade intrínseca do discurso, e não dos interlocutores. Trata-se de uma estrutura dinâmica por natureza, que registra as entidades, propriedades e relações mais proeminentes em cada enunciado ou segmento do discurso.[77] Essa informação é necessária para que o modelo possa refletir o fato de que os participantes de um evento comunicativo, além de estarem "focalizando" o que está sendo dito, estão também "focalizando" os motivos que os levam a dizer o que estão dizendo naquele momento do discurso.[78]

O esquema a seguir sugere uma possibilidade de integração de um módulo responsável pela representação abstrata do foco de atenção e das intenções dos participantes do discurso (as projeções δ):

---

77 Consideram-se proeminentes as entidades que tenham sido explicitamente mencionadas em um determinado segmento ou que passaram a ser salientes durante os processos de produção ou recepção dos enunciados contidos no segmento.

78 Para melhor compreensão do constructo *foco imediato*, remeto o leitor para a análise ilustrativa no final deste capítulo.



Esse esquema permite a visualização da teoria léxico-funcional, acrescida de informações semânticas, pragmáticas e discursivas. Com esses recursos, esse modelo talvez seja capaz de fornecer contructos que poderão formar uma base de representações integradas.

Uma análise ilustrativa

Considere mais uma vez o "pesadelo de Joaquim Cruz":[79]

[1] Repórter - *"Como você se sente agora?"*

Joaquim Cruz - *''Me livrei de um sonho ruim, recorrente e sugestivo: nele, eu não sabia mais* ***o horário da prova****,*
perguntava $\varnothing_1$ *a um colega*
que também não sabia $\varnothing_2$
e chegava atrasado na pista.''
(VEJA, 22/07/92: p.9)

Para responder à pergunta (2) – *O que o atleta pergunta ao colega ?* – além de "saber" que #Joaquim Cruz45# é o referente das expressões *Joaquim Cruz* e *o atleta*, que #colega32# é o referente das expressões *um colega* e *o colega* e que a informação solicitada encontra-se codificada como objeto direto do verbo *perguntar*, um SPLN precisa "saber" que, no português, essa função gramatical pode não estar realizada foneticamente e que, mesmo assim, existem meios de se estabelecer a conexão entre esse elemento nulo e seu antecedente.

O mesmo fenômeno ocorre em enunciados como:

[2] Sonia Braga - *''Todo mundo falou de-* ***o meu cabelo grisalho****.*
Eu pintei $\varnothing$*,*
mas não deu certo.''
FOLHA DE SÃO PAULO, 24/02/95: 5-10)

---

79 Em cada exemplo, o elemento grifado é o sintagma nominal do qual o *objeto nulo* **Ø** deriva sua interpretação.

[3] Repórter - *‘‘Há quanto tempo vocês vendiam* ***<u>drogas</u>****?’’*
Entrevistado - *‘‘Vendíamos* **Ø**
havia pelo menos dois anos.’’
(FOLHA DE SÃO PAULO, 23/03/95: 3-3)

[4] Repórter - *‘‘Vocês também usavam* ***<u>drogas</u>****?’’*
Entrevistado - *‘‘Não,*
nenhum dos meus irmãos usavam **Ø**.*’’*
(FOLHA DE SÃO PAULO, 23/03/95: 3-3)

[5] Repórter - *“Qual sua visão sobre* ***drogas*** *? ”*
Herbert Viana - *“Eu usei* **Ø**
muito pouco, porque elas servem pouco para mim.”
(VEJA, 05/07/95: p.9)

Como explicar essa característica do português? Como os interlocutores sabem que, em [1], por exemplo, os objetos nulos $\mathbf{Ø_1}$ e $\mathbf{Ø_2}$ e a expressão *o horário da prova* apontam para o mesmo referente específico e definido, i.e., “um horário específico e definido de uma prova específica e definida num tempo e espaço específicos e definidos”? Como criar um modelo computacional capaz de identificar e interpretar o referente do objeto nulo? A qual teoria lingüística recorrer?

Ao buscar respostas para essas questões, pretendo mostrar um possível encaminhamento para o tratamento computacional do objeto nulo que possa também contribuir para avaliar a adequação de modelos de descrição de análise lingüística. Com isso, procuro ilustrar o modo de

pesquisa que venho defendendo no âmbito do PLN: pesquisas que busquem seus fundamentos em teorias lingüísticas e computacionais visando à construção de teorias integradas.

É fato que o português, como outras línguas românicas, apresenta sujeito nulo em orações finitas, fato que tem sido explicado em função do sistema de morfemas flexionais de concordância número-pessoal entre o sujeito e o verbo da oração.[80]

Lobato (1986: 433) apresenta a seguinte explicação:

> "O que ocorre é que todas essas línguas têm uma morfologia verbal muito rica, que torna dispensável a manifestação fonética do pronome sujeito. O pronome sujeito e a flexão verbal seriam, então, redundantes entre si nessas línguas chamadas de línguas pro-drop ou línguas com sujeito nulo (ou oculto) exatamente por sua característica de permitir a não-manifestação fonética do sujeito."[81]

Essa explicação associa o fato de o pronome poder ser omitido, na posição de sujeito, nas orações finitas, em línguas como o português, o espanhol e o italiano, com o fato dessas línguas possuírem uma "morfologia verbal muito rica", responsável pelo fenômeno da concordância número-pessoal entre o verbo e o sujeito e, conseqüentemente, portador de informação suficiente para a recuperação do conteúdo relevante para a interpretação do sujeito nulo. Esse

---

80 O termo "sujeito nulo" é empregado para designar o que a gramática tradicional costuma chamar "sujeito oculto". O termo "nulo" é empregado para qualificar elementos lingüísticos gramaticalmente relevantes que não apresentam realização fonética.

81 Essa explicação é decorrência da Hipótese de Identificação proposta por Jaeggli (1982). Essa hipótese estabelece que um pronome poderá ser omitido se sua referência puder ser recuperada a partir de outros elementos explícitos na oração como, por exemplo, as marcas de concordância sujeito-verbo. Jaeggli & Safir (1989) apresentam uma resenha detalhada de problemas empíricos e teóricos específicos que a existência do sujeito nulo coloca para a teoria lingüística.

argumento, plausível para explicar a existência do sujeito nulo, de nada vale para explicar a existência do objeto nulo, uma vez que, pelo menos em línguas semelhantes ao português, não há qualquer marca de concordância entre o verbo e o seu objeto.

A caracterização lingüística do objeto nulo decompõe-se na tarefa de resolução de três subproblemas:[82]

Subproblema 1.1: contextualização do problema específico do objeto nulo no âmbito de fenômenos lingüísticos mais gerais e descrição de suas características básicas;

Subproblema 1.2: caracterização do objeto nulo à luz de uma proposta teórica por meio da qual seja possível determinar suas características gramaticais, bem como a construção de uma representação formal;

Subproblema 1.3: identificação de seu papel no discurso, uma vez que os contextos lingüístico e extralingüístico desempenham papéis determinantes no processo de sua interpretação.

Subproblema 1.1: contextualização do problema

Nota-se, em primeiro lugar, que o fenômeno do objeto nulo é parte de um fenômeno mais geral, o fenômeno da *elipse* (ou *substituição zero*), entendido como um processo de apagamento de quaisquer segmentos de enunciados identificáveis pelos contextos lingüístico e/ou não-lingüístico.[83] Esses segmentos podem ser das mais variadas categorias e ter dimensões sintagmáticas diversas: itens

---

82 Observo que não pretendo aqui desenvolver uma análise exaustiva do fenômeno do objeto nulo em português, análise que, pela sua complexidade, mereceria um trabalho à parte. A discussão do objeto nulo limita-se às ocorrências ilustradas em [1]-[5].

83 *Cf*. Halliday & Hasan (*op. cit.*: cap. 4), Greimas & Courtés (*op. cit.*: 140), Dubois *et al.* (*op. cit.*: 207) e Borba (1984: 190-200),

lexicais, constituintes ou adjuntos oracionais, partes de constituintes ou de adjuntos oracionais e até orações inteiras. Empregando a terminologia da gramática tradicional, o *objeto nulo* é concebido como um caso particular de elipse: a elipse do complemento direto ou indireto de verbos transitivos, complemento que se realiza sintaticamente por meio de sintagmas nominais, sintagmas preposicionais ou por orações substantivas objetivas. Em particular, só estou considerando a elipse de sintagmas nominais.

Do ponto de vista semântico, o objeto nulo pode ser um elemento *endofórico* ou *exofórico*, dependendo do modo de interpretação:[84]

- se sua interpretação puder ser determinada exclusivamente em função da interpretação de outra expressão lingüística, presente no enunciado (também conhecida como "antecedente"), trata-se de um elemento de natureza endofórica, ou elemento de "interpretação dependente do texto" (KAMEYAMA, 1985);
- se, por outro lado, sua interpretação não puder ser assim determinada, então se trata de um elemento de natureza exofórica, isto é, elemento de "interpretação independente do texto", caso em que o objeto nulo refere-se a alguma entidade concreta ou abstrata, saliente no dado contexto não-lingüístico.

Kameyama (*op. cit.*) propõe a seguinte tipologia de referência exofórica: ***dêitica***: quando Ø referir-se a alguma entidade pertencente ao contexto espaço-temporal em que o locutor se encontra como, por exemplo: "*Pode pegar* ***Ø***"; ***indicial***: quando Ø referir-se aos participantes do evento comunicativo, e.g.: "*Quer ajudar* ***Ø***?";

84 Os termos *elipse*, *substituição zero*, *endofórico*, e outros termos relacionados, *exofórico*, *anáforico* e *catafórico* são tomados de Haliday & Hasan (*op. cit.*).

arbitrária: quando Ø referir-se a pessoas ou coisas em geral como, por exemplo: "*Aquele muro não permite* **Ø** *ver o jardim*". Com esse tipo de referência exofórica, fica definitivamente descartada a possibilidade de construção de um modelo computacional capaz de interpretar o **Ø**. Não é possível exigir-se da máquina uma capacidade de interpretação que nem os intelocutores humanos possuem. No máximo, seria possível criar um modelo computacional que identificasse esses casos e fizesse as seguintes perguntas: Pode pegar o quê? quem?, Quer ajudar o quê? quem?, Aquele muro não permite a quem ver o jardim? a ninguém?

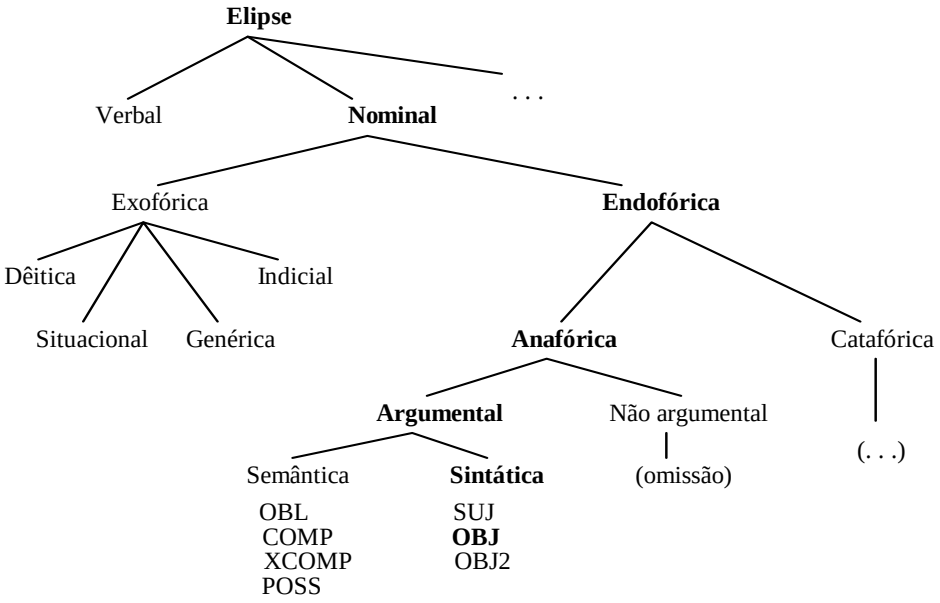
Resta, ainda, fazer uma última distinção, a que toma por base as posições sintagmáticas do objeto nulo em relação a seu antecedente. Quando o objeto nulo precede seu antecedente na cadeia sintagmática, diz-se que se trata de um elemento catafórico (aponta para frente); quando ocorre o inverso, tem-se um elemento anafórico (aponta para trás).

Essas considerações já são suficientes para a contextualização e caracterização geral do tipo de fenômeno lingüístico observado nos enunciados de [1] a [5]: trata-se de um fenômeno que se insere no domínio da elipse de sintagmas nominais que desempenham a função sintática de objeto. O conteúdo semântico do sintagma nominal elidido é recuperável por meio da relação anafórica que ele mantém com seu antecedente, localizado no contexto discursivo que o antecede.[85]

O esquema a seguir resume os tipos de elipse:

---

85 Kameyama propõe uma distinção entre anáfora sintática nula e anáfora semântica nula: a primeira ocorre somente nas funções SUJ e OBJ; a segunda, nas funções OBL e COMP. No caso da função ADJUNTO, não se tem anáfora, mas simplesmente uma omissão.



Subproblema 1.2: caracterização gramatical do *objeto nulo*

Identificar o estatuto gramatical do *objeto nulo* e propor mecanismos para sua interpretação são tarefas que exigem a adoção de um referencial teórico que forneça os elementos descritivos e interpretativos necessários e adequados ao tipo de análise que se pretende realizar.

Em estudo anterior (DIAS-DA-SILVA, 1992), fiz uma comparação entre as propostas de análise do *objeto nulo* desenvolvidas, de um lado, pela TPP e, de outro, pela TLF. A TPP caracteriza o *objeto nulo* como a instância de uma *variável*, ligada ao um *operador nulo*, coindexado com o *tópico do discurso*. Já a TLF caracteriza-o como uma instância de um *pronome nulo*. Como decidir entre essas propostas? **86**

Em um primeiro momento, descartei a análise proposta pela TPP, que postula que o *objeto nulo* é uma instância de uma *variável* (*cf.* RAPOSO, 1992). Tal análise é motivada pela análise do deslocamento de constituintes do tipo *quem*, em orações como ***Quem*** *você viu* $\mathbf{v_i}$*?*, que,

86 Além da motivação computacional, essa disputa teórica também contribuiu para a escolha desse fenômeno em particular.

por sua vez, espelha a relação que se estabelece entre um *operador* e uma *variável* no âmbito da lógica de predicados.[87] A teoria postula que todo elemento que se encontra deslocado de sua posição original deixa, nessa posição inicial, um *vestígio*, i.e., uma "cópia" de si mesmo sem conteúdo fonético.[88] No caso particular do deslocamento de constituintes do tipo *quem*, a teoria estipula que o deslocamento se dá a partir de uma posição argumental (neste exemplo, complemento do verbo *ver*) para uma posição não-argumental (posição de Especificador de C").[89] Os deslocamentos que envolvem elementos desse tipo, também denominados *constituintes Q*, são representados por meio dos dois constructos da lógica de predicados: no exemplo, ***quem*** é o *operador* e $\mathbf{v}_i$ a *variável*, um tipo particular de vestígio. A estrutura resultante do deslocamento é denominada estrutura-S.[90]

Diante dessa constatação, com base em Kameyama (*op. cit.*), que argumenta a favor da TLF para fenômeno semelhante no japonês, investiguei dados empíricos e propostas teóricas para a caracterização lingüística do *objeto nulo* no português do Brasil na dimensão da gramática. Em consonância com a proposta de Kameyama e

---

87 Haegeman (*op. cit.*: 443) esclarece a utilização dos termos *operador* e *variável* na TPP. Em expressões da lógica do tipo **∀x (Hx → Epx)**, isto é, "para todo x, se x é humano (H), então Pedro (p) enxerga (E)", ∀ é denominado operador e x uma variável ligada a ∀. Diz-se que a variável x está ligada a ∀ porque sua interpretação depende da interpretação de ∀.

88 Isto é, a posição por ele ocupada na estrutura-P – estrutura subjacente, projetada a partir do Léxico e das configurações sintagmáticas características da língua, em que se encontram representadas as relações temáticas e as dependências estruturais canônicas da língua.

89 As posições estruturais de sujeito de oração e de complemento são denominadas posições argumentais, ou posições-A. As demais posições são denominadas não-argumentais, ou posições não-A.

90 Por que, afinal, o *objeto nulo* é uma instância de uma *variável*? Conforme explica Raposo (*op. cit.*: 340) as orações que contêm objeto nulo devem necessariamente conter um "*tópico foneticamente nulo*, [um operador], cujo valor referencial é dado pelo contexto discursivo, ligando uma variável argumental na posição de objeto directo."

em oposição a modelos estritamente sintáticos como a TPP, mostrei que o *objeto nulo* no português do Brasil é uma instância da categoria pronominal nula *pro*: *pro objeto*.

Kameyama propõe uma tipologia lingüística em termos de dois parâmetros relacionados com as possibilidade de ocorrência da *anáfora zero* nas línguas naturais: *permissibilidade da anáfora zero* (PAZ) e *exigência de explicitação sintática* (EES).[91] Esses parâmetros podem ser interpretados como duas "forças" opostas e co-variantes que regulam o aparecimento da *anáfora zero* nas línguas naturais: (i) o parâmetro PAZ representa a *força pragmática*, que procura reduzir as redundâncias ao mínimo, permitindo que as funções gramaticais principais, SUJ e OBJ, sejam foneticamente nulas em orações finitas, quando seus referentes são facilmente recuperáveis no contexto discursivo, caso, portanto, da *anáfora zero*; (ii) a EES representa a *força sintática*, que exige o inverso: que essas mesmas FGs sejam morfologicamente explicitadas, mesmo quando codifiquem informações redundantes.

Há duas maneiras de codificação explícita das FGs: por elementos pronominais ou morfemas presos. Esses dois modos de codificação induzem uma classificação em termos de duas dimensões: uma dimensão estabelece a obrigatoriedade de codificação das FGs em termos de pronomes plenos e a outra, em termos de morfemas presos:

---

91 Entende-se por *anáfora zero*, expressões pronominais foneticamente vazias, cujo referente é determinado exclusivamente pela relação de co-referência que se estabelece entre esse elemento nulo e uma outra expressão (seu antecedente) já presente no contexto lingúístico.

		Codificação obrigatória de morfemas presos	
		SIM	NÃO
Codificação obrigatória de pronomes plenos	SIM	Tipo I	Tipo IV
	NÃO	Tipo II	Tipo III

A caracterização e exemplificação de cada um dos tipos são dadas a seguir.

Tipo I: línguas que tendem a codificar as FGs em termos de pronomes plenos e morfemas presos; por exemplo, inglês, alemão e francês;

Tipo II: línguas que tendem a codificar as FGs em termos de morfemas presos, mas não em termos de pronomes plenos; por exemplo, espanhol, italiano, turco e português europeu;

Tipo III: línguas que tendem a codificar as FGs em termos de pronomes foneticamente vazios; por exemplo, japonês, chinês, coreano e português europeu;[92]

Tipo IV: línguas que tendem a codificar as FGs em termos de pronomes plenos, mas não em termos de morfemas presos; por exemplo, holandês, norueguês e sueco.

Como o português do Brasil se encaixa nessa tipologia? O português do Brasil apresenta características sintáticas peculiares que o diferenciam tanto do português europeu como das demais línguas românicas. Em particular, apresenta-se como uma língua em que fatos discursivos indiscutivelmente interagem e se sobrepõem à gramática (*cf.* RAPOSO, 1986).

92 Kameyama argumenta que o português europeu é do Tipo II em relação ao SUJEITO e do Tipo III em relação ao OBJETO, o que para ela significa que o português europeu encontra-se em uma fase de transição entre esses dois tipos.

Embora as argumentações centrem-se em fatos eminentemente sintáticos, delas é possível depreender alguns traços característicos do português do Brasil, relevantes para a discussão. Entre elas os autores têm apontado que o português do Brasil, como o português europeu, possui a característica de ser uma língua "orientada para o discurso" e o *objeto nulo*:

- refere-se a uma expressão contida no contexto lingüístico precedente (anáfora);
- distingue-se do objeto nulo das demais línguas românicas, incluindo o português europeu, porque pode ter ambos os tipos de referência [+/- definida];
- é um *pro objeto* gerado na estrutura-P que pode ter os seguintes traços [+ referencial], [+/- humano], [+/- definido], [+/- endofórico], [+/- animado].

Além dessas, apontam também que o português do Brasil, em contraste com o português europeu, apresenta algumas propriedades tipológicas singulares:

- o pronome foneticamente vazio *pro* pode ter referência [+ definida];
- o morfema de concordância verbal parece ter perdido seu papel de identificador do sujeito nulo;[93]
- o tópico frasal é re-analisado como sujeito;
- os clíticos são raramente encontrados na modalidade oral coloquial.

93 Fato que parece confirmar a previsão de Kameyama de que o tópico frasal é colocado em uma posição superior ao sujeito na hierarquia que estabelece a ordem de preferência dos elementos controladores da correferência dos pronomes nulos.

Essas características são apontadas por Kameyama como indicadores de que a língua segue o parâmetro PAZ. Além disso, como os elementos de "contéudo fonético mínimo", isto é, os clíticos, candidatos a codificar informação recuperável pelo contexto do discurso ou pelo contexto situacional, deixaram de ser utilizados pelos falantes em relação à FG OBJ, parece plausível concluir que o português do Brasil segue o parâmetro PAZ, isto é, o *objeto nulo* é um pronome foneticamente vazio.

Observo que Kameyama argumenta que as línguas, em geral, codificam as informações facilmente recuperáveis pelos contextos discursivo e/ou situacional por meio de elementos pronominais com conteúdo fonético mínimo. No caso do português, esses elementos seriam os clíticos.

Lembre-se de que, na TLF, os elementos sintáticos com conteúdo fonético nulo não são representados na estrutura-c. Em outras palavras, a estrutura sintagmática da frase não possui nós vazios, como na teoria chomskiana da regência e ligação. Logo, nessa estrutura, ele simplesmente não é representado. Sua **identificação** é possível por meio da estrutura de argumentos do predicador e sua **representação**, na estrutura-f, é o pronome anafórico *pro* que preenche o valor da função OBJ.

O processo de identificação é o seguinte:

Para cada predicador detectado no texto,

1. especificar sua estrutura de argumentos;

2. verificar se todas as funções gramaticais obrigatórias estão preenchidas por conteúdo lexical, caso contrário, marcar as funções não preenchidas como omissões;

3. rotular a omissão que corresponde à função OBJ como *pro objeto*.

Uma vez representado e identificado gramaticalmente, resta agora, representar o *objeto nulo* em um modelo que inclua informações sobre o discurso.

Subproblema 1.3: identificação do papel do objeto nulo no discurso

Retomando a idéia do foco imediato de atenção inicialmente proposto por Grosz e Sidner, Kameyama desenvolve uma "teoria da centralização" que visa a explicar as condições de boa formação das conexões mais locais que se estabelecem entre os enunciados. Postula-se que cada enunciado pode estar associado a nenhum ou a uma variedade de centros potenciais, um dos quais é o CENTRO, ou seja, a entidade "mais saliente" do contexto situacional, focalizada pelos participantes do discurso e lingüisticamente codificada em um sintagma nominal referencial. O CENTRO é concebido como um conceito mental que ocupa a atenção do falante e/ou ouvinte a cada "momento" do processo comunicativo – é o foco de atenção da EA.[94]

94 A EA pode ser concebida como um "caderno de anotações" em que, a cada momento, registram-se as entidades que estão em foco.

Assim, a teoria estipula que o CENTRO (caso exista um), para cada segmento do discurso, é único e o segmento de discurso que o contém é denominado “unidade de centralização”, que representa a menor unidade discursiva. Isso quer dizer que, à medida que o discurso se desdobra, os participantes do evento comunicativo, a cada momento, focalizam sua atenção em uma única entidade por vez. Logo, o principal papel discursivo do CENTRO é estabelecer a conexão entre os enunciados que compõem uma unidade de centralização. Como o CENTRO, por definição, refere-se a uma entidade anafórica, Kameyama postula que:

(i) o enunciado inicial do discurso não está associado a CENTRO algum, mas pode estar associado a centros potenciais;

(ii) cada unidade de centralização termina quando no enunciado subseqüente, o foco de atenção desloca-se para outra entidade. Em outras palavras, o CENTRO deixa de ser o CENTRO do enunciado seguinte, que passa novamente a conter centros potenciais.

Um exemplo típico de unidades centralizadas é ilustrado a seguir, em que 1-8 são enunciados; 1-3, 4-5 e 6-8 são unidades de centralização e cada par C2 = C3 e C7 = C8 são os centros potenciais das unidades:

unidade	c11,c21,…,cn1	
de	C2,c12,…,cn2	
centralização-1	C3,c13,…,cn3	C2=C3
unidade de	c14,c24,…,cn4	
centralização-2	C5,c15,…,cn5	

unidade	c16,c26,…cn6	
de	C7,c17,…,cn7	
centralização-3	C8,c18,…cn8	C7=C8

Tal modelo representa um processo cognitivo hipotético que regula o processamento do discurso. Um exemplo concreto que realiza essas unidades é dado pelo segmento de diálogo hipotético a seguir:

unidade	— Comprei uma casa (c11)
de	— Quanto você pagou? C2=CASA
centralização	— Muito barato. R$ 30 mil.C3=CASA (c13) **95**
unidade de	— Você precisou vender o carro? (c14)
centralização (c15)	— Não precisei vender. Eu tinha o dinheiro .C5=CARRO
unidade	— A casa está em boas condições? (c16)
de	— Está boa.Só precisa de uma pintura. C7=CASA (c17)
centralização	— A minha também está precisando. C8=CASA

Segundo Kameyama, os pronomes anafóricos foneticamente vazios podem ser marcas do CENTRO de um segmento de discurso. Decorre, então, que um dos centros potenciais, quando passam a CENTRO tornam-se um pronome nulo. Assim, quando o *pro objeto* é o CENTRO, sua função discursiva é sinalizar o foco de atenção dos participantes do discurso.

O processo de identificação e retenção do CENTRO, para o processamento computacional do *objeto nulo*, é regulado por meio de duas regras heurísticas:

95 Observe que o foco é o preço da casa, mas o CENTRO continua sendo CASA.

- **Regra de Estabelecimento do CENTRO (E)**: quando um dos centros potenciais do enunciado anterior torna-se o CENTRO do enunciado subseqüente, um pronome anafórico foneticamente vazio deverá ser utilizado. Se mais de um pronome com essa característica for detectado, aquele que preencher a função hierarquicamente superior será o CENTRO. A hierarquia de funções em ordem decrescente é dada por:
  SUJ TOPICALIZADO > OBJ TOPICALIZADO > POSS TOPICALIZADO .> SUJ > OBJ > POSS > outras funções;
- **Regra de Retenção do Centro (R)**: essa regra é semelhante à anterior. A diferença reside nas restrições que ela impõe sobre as condições a que o CENTRO deve obedecer para que possa ser mantido. O CENTRO é mantido se o pronome nulo do enunciado seguinte preencher a mesma função gramatical.

Assim, o processo de identificação e interpretação do *objeto nulo* segue os seguintes passos:

1. especificar a estrutura de argumentos do predicador de cada frase que integra os enunciados;

2. verificar se todas as funções gramaticais obrigatórias estão preenchidas por conteúdo lexical, caso contrário, marcar as funções não preenchidas como omissões;

3. rotular a omissão que corresponde à função OBJ como *pro objeto*;

4. aplicar as regras E e R.

Para ilustrar o modelo em funcionamento tomemos o enunciado [1], que, para fins de exemplificação, foi reduzido aos seus elementos essenciais:

me livrei de <u>um sonho</u> (centro potencial 11= **OBL** = **SONHO**)

PRED = ‘livrar < (SUJ=1s nom) (OBJ=1s acus) (OBL=UM SONHO) >’

nele, eu não sabia <u>o horário-da-prova</u> (centros potenciais 21 = **OBJ** = **HORÁRIO DA PROVA** e 22 = 11 = **OBL** = **SONHO**)

PRED = ‘saber < (SUJ=1s nom) (OBJ=HORÁRIO DA PROVA) >’

perguntava *pro* a <u>um colega</u> (***pro*** $\Rightarrow$ **CENTRO 3** = **OBJ** = **HORÁRIO DA PROVA** e centro potencial 31 = **OBL** = **COLEGA**)

PRED = ‘perguntar < (SUJ=1s nom) (OBJ=*pro*) (OBL=UM COLEGA) >’

que também não sabia *pro* (***pro*** $\Rightarrow$ **CENTRO 3** = **CENTRO 4** = **OBJ** = **HORÁRIO DA PROVA** e centro potencial 41 = 31 = **SUJ** = **COLEGA**

PRED = ‘saber < (SUJ=3s nom) (OBJ=*pro*) >’

e chegava atrasado na pista

(centro potencial 51 = 31 = **OBL** = **PISTA**

PRED = ‘saber < (SUJ=3s nom) (OBJ=*pro*) >’

Uma avaliação dessa abordagem do *objeto nulo*, no português, parece bastante útil não apenas para se ter uma compreensão melhor da dinâmica da pesquisa lingüística e de sua contribuição para o PLN, como também reforçar a argumentação de que abordagens que integrem as faces gramatical e discursiva dos fenômenos da linguagem são mais adequadas do que as estritamente sintáticas (*cf*. BRESNAN & KANERVA, 1988). Evidentemente, tal tarefa está longe de ser fácil.

CAPÍTULO 5 – Equacionamento do domínio representacional

> *"The central problem in natural language processing is the translation of the potentially ambiguous natural-language input into an unambiguous internal representation, i.e., internal to the program doing the processing."*
>
> Jaime G. Carbonell & Philip J. Hayes (1990: 662)

A caracterização da análise gramatical é fundamental para o desenvolvimento de SPLNs. Todavia, tanto a concepção quanto a finalidade desse procedimento de análise, cujas origens remontam à tradição gramatical do ocidente, têm sofrido modificações em virtude de sua relevância para uma série de disciplinas, incluindo-se entre elas: a lingüística teórica, a psicolingüística, a teoria das linguagens formais, a ciência da computação e a inteligência artificial.

O quadro, abaixo, construído a partir das esclarecedoras discussões de Karttunen & Zwicky (1985), mostra as modificações e explicita o conceito e o propósito do procedimento de análise gramatical, relevantes para o PLN:

	Gramática Tradicional	**Teoria Lingüística**	**Teoria das Linguagens Formais**	**Linguagens de Programação**	**PLN**
Agente da operação	seres humanos	dispositivos abstratos	dispositivos abstratos	computadores	computadores
Objeto da operação	representação ortográfica de frases	representação abstrata de frases e textos em uma língua natural	seqüência de qualquer tipo de símbolos	representação abstrata de "frases" e "textos" em uma língua artificial	representação abstrata da forma ortográfica de um subconjunto de frases e textos em uma língua natural
Resultado da operação	descrições informais, que,em si, constituem um discurso lingüístico, composto de frases de alguma língua natural, cujo léxico inclui os termos técnicos empregados nas descrições	descrições precisas, contendo objetos lingüísticos formais: categorias, estruturas e relações	descrições precisas, contendo objetos formais: categorias, estruturas e relações	descrições precisas, contendo representações de objetos formais: categorias, estruturas e relações, cuja interpretação resulta em mudanças de estados internos da máquina	escrições precisas, contendo representações de objetos lingüísticos formais: categorias, estruturas e relações, cuja nterpretação resulta em mudanças de estados internos da máquina
Natureza da operação	heurística	algorítmica e heurística	algorítmica	algorítmica ou heurística	algorítmica e heurística
Finalidade da operação	didático-normativa	teórico-descritiva	teórico-descritiva	descritivo-operacional	descritivo-operacional
Aspectos privilegia-dos pela operação	aspectos estritamente sintáticos	aspectos sintáticos e aspectos semântico-pragmáticos sintaticamente relevantes	aspectos estritamente sintáticos	aspectos sintáticos, semânticos e pragmáticos	spectos sintáticos, semânticos e pragmáticos
Modo de aplicação da operação	consciente e seu domínio varia de indivíduo para indivíduo	irrelevante	irrelevante	automático	automático
Aquisição	adquirida por treinamento específico ou prática explícita	controvertida	irrelevante	irrelevante	irrelevante

Dias da Silva, B.C. (1996)

As modificações, que ocorreram com a passagem da concepção tradicional para a concepção introduzida pela teoria lingüística, somadas às novas concepções introduzidas a partir desta pela teoria das linguagens formais e linguagens de programação, são cruciais para delimitar o conceito de processamento gramatical no âmbito do PLN.

Na passagem da gramática tradicional para a teoria lingüística, destacam-se as mudanças que ocorreram em relação ao **agente** do processo de análise (que passa a ser qualquer dispositivo abstrato), ao seu **objeto** (que passa a ser representações formais), ao seu **resultado** (que passa ser objetos lingüísticos formais), à sua **natureza** (que também pode ser considerada algorítmica) e, finalmente, ao seu **propósito** (que deixa de ser um "exercício" para aprimorar a utilização dos meios de expressão disponíveis nas línguas e passa a ser uma atividade teórico-descritiva) e à sua **ênfase** (que passa a incluir elementos semânticos e pragmáticos).

Na passagem da teoria lingüística para o PLN, intermediada pela teoria das linguagens formais e das linguagens de programação, destacam-se as alterações que dizem respeito ao **agente** da operação (que agora concretiza-se no próprio computador), ao seu **objeto** (que passa a ser uma representação de uma representação), ao seu **resultado** (que passa a ser a representação de objetos lingüísticos que fornece os elementos para a simulação do comportamento lingüístico humano), ao seu **propósito** (que passa a ser descritivo-operacional), à sua **ênfase** (que se estende a todos os níveis de análise) e ao **modo** de aplicação (que é necessariamente automático).

Esse redimensionamento do processo de análise gramatical é, sem dúvida, resultado de mudanças da própria maneira de se conceber uma gramática, que passa a ser definida por um conjunto de princípios do tipo S = SN + SV.

Uma gramática formal, que passa a conter um sistema de princípios dessa natureza, pode ser "transformada" em um conjunto de regras que analisam, geram ou simplesmente verificam estruturas gramaticais bem-formadas. Assim, uma gramática pode assumir três funções bastante distintas: a função de análise estrutural, a de produção de estruturas e a de verificação de estruturas.

De fato, as teorias lingüísticas contemporâneas têm enfatizado não mais a especificação de regras, mas a especificação de um conjunto de condições de boa-formação que devem ser aplicadas durante o processamento das frases. As gramáticas propostas por Chomsky (1981 e 1992), Gazdar (*op. cit.*) e Bresnan (1982), por exemplo, não privilegiam conjuntos de regras e não são essencialmente derivacionais, no sentido de derivação proposto em Chomsky (1965). Todas elas estabelecem sistemas de restrições que regulam e sancionam as possíveis estruturas gramaticais admissíveis por uma determinada língua.

Com efeito, embora uma gramática e os procedimentos de análise sintática sejam estreitamente relacionados, é importante observar que ambos são objetos conceitualmente distintos. A gramática passa a ser concebida como uma definição abstrata de um conjunto de objetos estruturados, isto é, é a especificação das categorias sintáticas, dos itens lexicais, da forma dos objetos lingüísticos e das regras e princípios de estruturação desses objetos. Os procedimentos de análise gramatical, por

sua vez, são concebidos como algoritmos que, valendo-se da gramática, executam a construção desses objetos. Essa distinção é decisiva, pois permite separar o estudo dos modos de representação das gramáticas (teorias da competência) do estudo dos procedimentos responsáveis por seu funcionamento (teorias do desempenho).

Os dois dispositivos computacionais que simulam essas duas dimensões (competência e desempenho) são, respectivamente, as gramáticas sintagmáticas, simples ou ampliadas, e os analisadores gramaticais (do inglês *parsers*).**96** O primeiro dispositivo assume a forma de estruturas de dados que armazenam as regras e as estruturas gramaticais e o léxico. O segundo dispositivo assume a forma de algoritmos, cuja função é estabelecer a correspondência entre a forma ortográfica superficial de frases ou textos e suas respectivas estruturas abstratas. O resultado da operação dos analisadores gramaticais são descrições precisas, contendo representações dos objetos lingüísticos relevantes para o processamento semântico, bem como outras informações necessárias para o processamento pragmático-discursivo.

Assim, o equacionamento do domínio representacional do PLN envolve necessariamente a discussão de questões em três subdomínios:

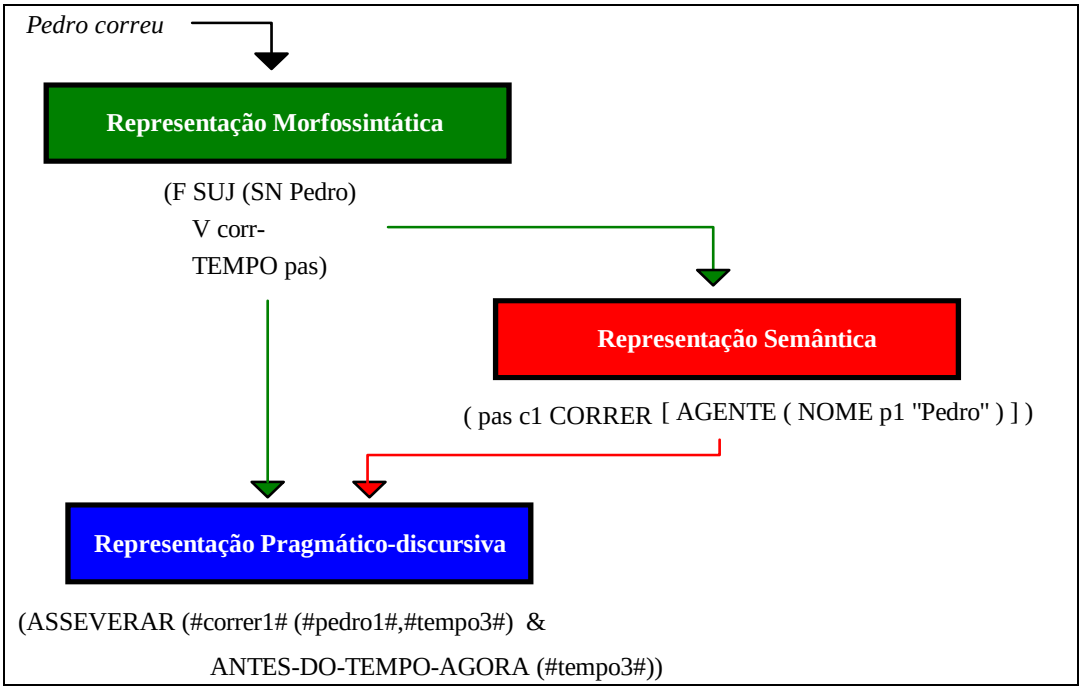
- morfossintático, que trata da **representação** das gramáticas e dos analisadores gramaticais, incluindo a representação das regras e das estruturas morfossintáticas e de léxicos enriquecidos com informações pragmático-discursivas;

---

96 O termo inglês *parsing*, por sua vez, deriva-se da expresão latina clássica *pars orationis*.

- semântico, que trata da **representação** de estruturas semânticas, de domínios conceituais e de estratégias computacionais de interpretação dessas representações;
- pragmático-discursivo, que trata da **representação** da estrutura do discurso e dos contextos pragmático-discursivo e situacional.

Baseando-me na ilustração de Allen (*op. cit.*: 18), retomo o exemplo do capítulo anterior (*Pedro correu*) para concretizar os níveis sucessivos de representação:



Representações dessa natureza precisam ser necessariamente explícitas, consistentes e, principalmente, não-ambíguas, para que possam ser transformadas em programas computacionais. Para isso, nada mais natural que se recorra a uma série

de formalismos, entre os quais a lógica clássica, a lógica de tipos, a lógica modal, o cálculo λ, os grafos, as matrizes de atributo-valor, os *frames* e as funções matemáticas. Esses formalismos, por sua vez, formam a base de representação para gramáticas, estruturas gramaticais, redes de transição, léxicos, analisadores gramaticais, redes semânticas, formas lógicas, estruturas semânticas, pragmático-discursivas e contextuais.

Há que se ressaltar que essa decomposição em três subdomínios complementares reflete, por um lado, a própria sistematização clássica da teoria lingüística (já delineada no capítulo anterior). Por outro, reflete também os três níveis do PLN (*cf.* ALLEN, *op. cit.*: 8):

(N1) nível de identificação, segmentação e estruturação do material lingüístico a ser processado;[97]

(N2) nível de interpretação semântica descontextualizada das estruturas construídas no nível anterior;[98]

(N3) nível de contextualização da interpretação semântica, ou nível de representação do significado final.

Esses três níveis são conceitualmente necessários. Sua ordenação, porém, é dependente do tipo de SPLN que se pretenda projetar. Mecanismos de controle em cada nível, por exemplo, poderão ser previstos para decidir quando transferir o processamento de um nível para o outro. A desvantagem de se manter processamentos estanques em

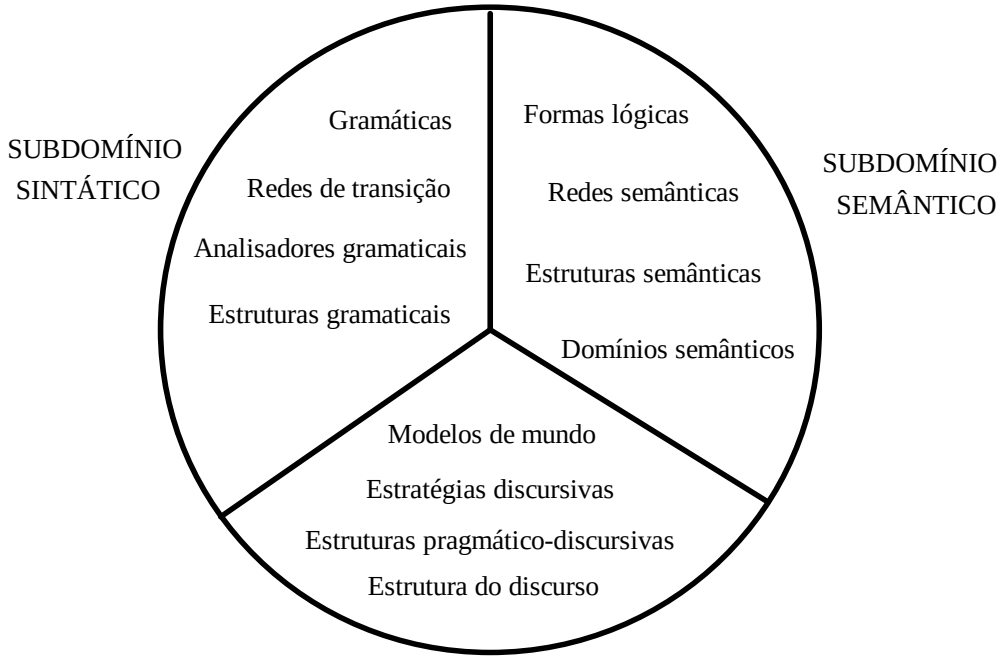
---

97 Por material lingüístico entendo o tipo de objeto lingüístico envolvido no processamento: morfemas, itens lexicais, sintagmas, frases, segmentos de texto e textos.

98 Moore (1981), voltando a atenção para a determinação do significado não contextualizado dos enunciados, argumenta a favor desse nível intermediário de representação semântica, distinto do nível discursivo em que o significado ‘‘pleno’’ do enunciado é determinado.

cada nível é evidente, considerando-se que a resolução dos vários tipos de ambigüidades e subespecificações, detectados em um determinado nível, depende de informações fornecidas pelo nível seguinte. Sendo assim, um SPLN, que mantenha esse tipo de estrutura terá seu desempenho comprometido, uma vez que uma frase poderá apresentar um número assustadoramente grande de estruturas sintáticas bem-formadas. Kurtzweil (*op. cit.*: 306) comenta que uma frase pode até mesmo apresentar um milhão de diferentes estruturas sintáticas gramaticalmente corretas!

O esquema a seguir sintetiza tudo isso:

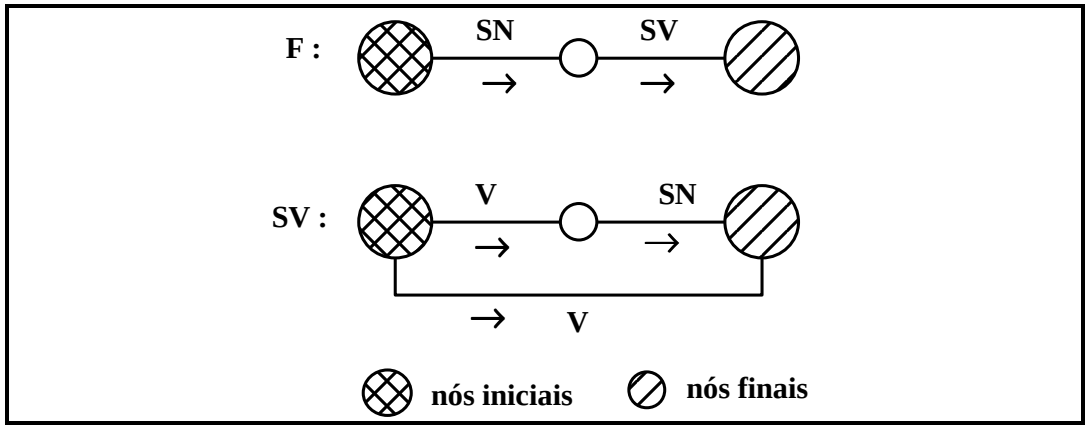


Subdomímio morfossintático

Gramáticas

Os dois formalismos básicos, empregados para representar gramáticas e léxicos, são os formalismos projetados para as **Redes de Transição Recursivas** (RTR)[99] e para as **Gramáticas Livres de Contexto** (GLC).[100] O esquema abaixo ilustra o primeiro tipo de formalismo:

Rede de Transição Recursiva (RTR)



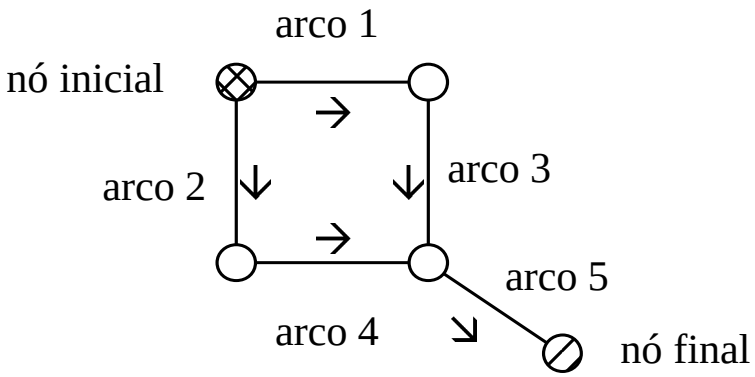
Símbolos terminais (Léxico) para esse formalismo:

SN:	Paulo Ana	**V**:	encontrou matou viu viajou

99 *Cf.* Woods (1970 e 1990); Sudkamp (*op. cit.*); Joshi (1990).

100 *Cf.* Chomsky (1957 e 1965); Kimbal (1973); Gazdar (*op. cit.*); Révész (1983); Joshi (*op. cit.*).

Cada rede que compõe uma RTR é representada como mostra a figura:



Como diz Gazdar, as RTRs podem ser interpretadas como mapas que nos auxiliam a percorrer as estruturas sintáticas das línguas. Para decidirmos se uma dada seqüência de palavras constitui uma estrutura sintática bem-formada, precisamos encontrar, na RTR, um caminho conexo que liga o nó inicial ao nó final.

Operacionalmente, procederíamos da seguinte maneira: colocamos um ponteiro indicador sobre o nó inicial da RTR (suponhamos que a rede seja a rede representada na figura acima) e outro ponteiro indicador sobre a primeira palavra da seqüência. Consultando o léxico, confrontamos o nome da categoria sintática da primeira palavra da seqüência com o(s) nome(s) da(s) categoria(s) do(s) arco(s) que parte(m) do nó inicial. Se, na RTR, existir pelo menos um arco com o mesmo nome, saltamos o ponteiro indicador que está sobre a rede para o nó seguinte e saltamos o ponteiro indicador que está sobre a seqüência de palavras para a palavra seguinte.

Procedemos de maneira análoga para as palavras subseqüentes. Se, em algum momento de nossa trajetória, duas situações especiais ocorrerem simultaneamente – o ponteiro indicador que está

sobre a rede encontrar-se sobre o nó final e o ponteiro indicador que está sobre a seqüência de palavras encontrar-se depois da última palavra da seqüência – então podemos concluir que obtivemos sucesso na construção do caminho procurado.

Caso não seja possível encontrar esse caminho, porque houve alguma etapa em que o confronto dos nomes (nome da categoria sintática da palavra e o nome do arco) revelou uma disparidade, ou porque não houve a simultaneidade das duas situações especiais, poderemos seguramente concluir que, segundo a gramática especificada pela RTR, a seqüência analisada não constitui uma estrutura sintática bem-formada.

As redes de transição recursivas são conjuntos desse tipo de rede. Cada uma delas é rotulada com o nome de um tipo específico de constituinte oracional: frase, sintagma nominal, sintagma verbal, sintagma preposicional, sintagma adverbial, sintagma adjetivo, especificador, sintagma preposicional, etc.. Sua topologia é constituída por uma seqüência de nós (estados), interligados por arcos (transições), rotulados com o nome de constituintes oracionais, ou com o nome de uma categoria sintática: substantivo, verbo, preposição, adjetivo, advérbio, determinante, modificador, conjunção, etc.. Além dessas características, cada rede possui um conjunto de nós que delimitam o seu início e fim, denominados, respectivamente, nós iniciais e finais.

As gramáticas livres de contexto, por sua vez, bastante conhecidas no âmbito da lingüística, possuem os seguintes constructos: símbolos categoriais, que especificam as categorias e estruturas sintáticas admissíveis, símbolos não-terminais, que especificam os itens

contidos no Léxico, e regras de reescrita, que especificam os tipos de estrutura sintática bem-formadas. Observe o exemplo:

Gramática Livre de Contexto (GLC)

Símbolos categoriais:	**Regras de produção:**
F (Frase)	F → SN SV
SN (Sintagma Nominal)	SV → V SN
SV (Sintagma Verbal)	SV → V
V (Verbo)	

Símbolos não-terminais (Léxico) para a gramática

SN:	Paulo	**V**:	encontrou
	Ana		matou
			viu
			viajou

Esses dois formalismos básicos (RTR e GLC) podem ser acrescidos de registros definidos para cada constituinte oracional e de dispositivos de teste e de procedimentos computacionais de manipulação desses registros. Os registros podem ser programados para, por exemplo, armazenar as informações sobre certos atributos da estrutura que está sendo construída e processada (por exemplo, SUJEITO, OBJETO, ADJUNTO, etc.), e os dispositivos servem para testá-los quanto à aceitabilidade de seus conteúdos.

Como esses formalismos contêm registros manipuláveis e têm a propriedade de transferir o controle do processo computacional,

em função dos estados em que os registros se encontram, essas redes ampliadas são equivalentes às máquinas de Turing. Logo, são capazes de representar qualquer tipo de gramática.[101] Outro modo de extensão desses formalismos básicos é acrescentar mecanismos de restrição e dispositivos para a armazenagem de constituintes que resultam do processo de unificação (KAY, 1986).[102] Nas gramáticas lógicas, esse dispositivo consiste em criar um argumento extra nas estruturas dos termos (ABRAMSON & DAHL, *op. cit.*; BRATKO, 1986; McCORD, 1990).

Com essas extensões, os dois formalismos básicos transformam-se em **Gramáticas Livres de Contexto Ampliadas** (GLCA) e nas **Redes de Transição Ampliadas** (RTA), propostas por Woods, na década de 70 (*cf.* WOODS, 1990). A gramática léxico-funcional é um tipo de GLCA em que as representações das regras sintagmáticas e dos itens lexicais de uma GLC foram ampliadas com o acréscimo dos esquemas funcionais, dos pares de atributo-valor e dos mecanismos de unificação de estruturas (BRESNAN, 1982).

Os "constituintes deslocados", amplamente estudados pela gramática gerativo-transformacional, podem ser representados e processados por meio desses formalismos. Os deslocamentos limitados (por exemplo, o deslocamento do objeto nas orações passivas) podem ser formalizados com a especificação de testes e procedimentos apropriados

---

101 As máquinas de Turing, formuladas pelo matemático britânico Alan Turing, são constructos matemáticos empregados no estudo das limitações dos sistemas computacionais. Essencialmente, elas representam a estrutura lógica de qualquer dispositivo computacional passível de ser construído. Em outras palavras, as máquinas de Turing podem ser consideradas a abstração matemática de um computador.

102 Kay (1985) propõe uma *gramática de unificação funcional*. Essa gramática utiliza a unificação como a operação básica no processo de construção da representação sintática da frase. No âmbito da lingüística computacional, cita-se o formalismo de representação gramatical *PATR* (*Pattern Recognition*) (*cf.* SHIEBER, 1986).

de manipulação de registros, ampliações já constantes dos formalismos GLCA e RTA. Já os deslocamentos ilimitados (por exemplo, o deslocamento de constituintes em orações topicalizadas, orações relativas e perguntas parciais) dependem de novas ampliações: acréscimo de listas de armazenagem que possibilitem a retenção de constituintes e de procedimentos que reutilizem esses constituintes retidos.

Relação entre línguas, gramáticas, redes de transição e autômatos

Dados um conjunto finito de elementos atômicos, um conjunto finito de símbolos categoriais e um conjunto finito de regras de produção, qualquer língua natural (L) pode ser concebida como um conjunto infinito de frases. Cada frase, por sua vez, pode ser concebida como uma seqüência finita de um ou mais itens lexicais que constituem o léxico de L. A gramática de uma língua é, então, definida como a especificação formal finita do conjunto infinito de frases pertencentes à L (PARTEE *et al., op. cit.*).

Do ponto de vista estritamente formal, uma gramática pode ser caracterizada em termos de três conjuntos disjuntos, um dos quais possui um (ou mais) símbolo(s) diferenciado(s), denominado(s) símbolo(s) inicial(is):

- um conjunto de **Símbolos Não-terminais** N;[103]
- um conjunto de **Símbolos Terminais** T, contendo o **Símbolo Inicial** F;
- um conjunto de **Regras de Produção** P, também chamadas regras de reescrita, da forma geral: $\alpha \rightarrow \beta$, em que α e β são seqüências de símbolos. α contém pelo menos um símbolo não-terminal e β resulta da substituição de algum símbolo não-terminal de α por um seqüência de símbolos vazia ou por uma seqüência de símbolos de T e de N.[104]

Essas especificações formais podem ser empregadas como modelo abstrato das línguas naturais, estabelecendo-se as seguintes correspondências:

- símbolos categoriais $\leftrightarrow$ categorias de classificação dos tipos sintáticos, por exemplo F (frases), N (nomes), V (verbos), A (adjetivos), P (Preposições), SN (sintagmas nominais), e assim por diante;
- regras de produção $\leftrightarrow$ regras sintáticas que determinam as estruturas sintáticas permissíveis, por exemplo F = SN SV;

---

103 Em uma gramática sintagmática, há dois tipos de símbolos: os símbolos terminais, isto é, os itens lexicais (lexemas e morfemas), e os símbolos não-terminais, que especificam as várias categorias sintáticas a que cada item lexical pertence. Os símbolos terminais são assim chamados, porque as regras de produção não se aplicam a eles.

104 As regras de produção especificam como os símbolos não-terminais são transformados para produzir as frases da língua. Durante o processo, aplicam-se regras de produção e regras de substituição lexical, nessa ordem. No final do processo, também conhecido como processo de derivação, todos os símbolos não-terminais são substituídos por símbolos terminais, isto é, itens lexicais.

- símbolos terminais ↔ itens lexicais, por exemplo, *casa*, *canta-*, *-r*, *de*, etc.

O estudo das linguagens formais mostra que a forma das *regras de produção*, mais conhecidas entre os lingüistas como *regras de reescrita*, determina o tipo de gramática e, conseqüentemente, o tipo de línguagem por ela gerada. Um dos resultados importantes desse estudo, que provocou calorosos debates sobre a capacidade gerativa da gramática transformacional, é a hierarquia que se estabelece entre os possíveis tipos de gramaticas formais, conhecida como *Hierarquia de Chomsky*. Essa hierarquia identifica quatro classes de gramáticas em função do tipo de regra:

Forma das regras do Tipo 0:

Não há restrição alguma em relação à forma das regras: $x \rightarrow y$, em que *x* é uma seqüência não nula de símbolos, contendo pelo menos um símbolo categorial, e y é uma seqüência nula ou uma seqüência de um ou mais símbolos categoriais e/ou terminais.

Forma das regras do Tipo 1:

A forma da regra é dada por: $\boldsymbol{x\ A\ z \rightarrow x\ y\ z}$, em que *A* é um símbolo não-terminal, *y* é uma seqüência de um ou mais símbolos, terminais ou não, e os símbolos *x* e *z* são nulos ou seqüências de um ou mais símbolos, categoriais e/ou terminais. Além disso, o número de símbolos à direita da regra deverá ser, no mínimo, igual ou superior ao número de símbolos à esquerda da regra.

Forma das regras do Tipo 2:

A forma da regra é dada por: $\boldsymbol{A \rightarrow x}$, em que *A* é um símbolo categorial e *x* é uma seqüência de um ou mais símbolos, categoriais e/ou terminais, podendo também ser uma seqüência nula.

Forma das regras do Tipo 3:

A forma da regra é dada por: $\boldsymbol{A \rightarrow x\ B}$ ou $\boldsymbol{A \rightarrow x}$, em que *A* é um símbolo categorial e *x* representa um símbolo terminal.

Segundo o grau de complexidade decrescente, identificam-se os quatro tipos de gramáticas:

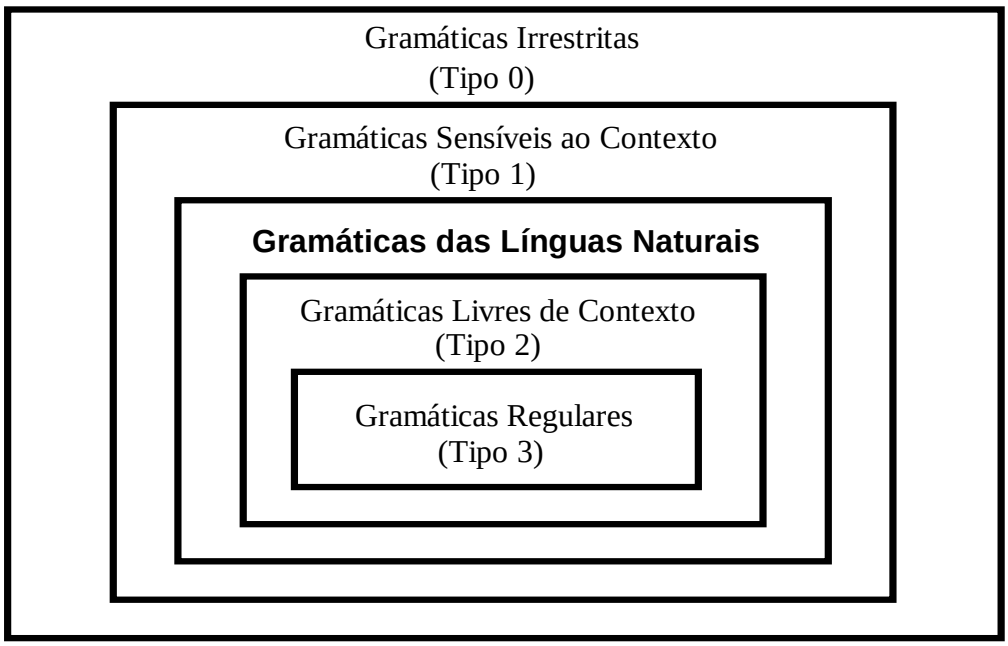
- irrestritas (tipo 0) geram línguas L0;
- sensíveis ao contexto (tipo 1) geram línguas L1;[105]
- livres de contexto (tipo 2) geram línguas L2;
- regulares (tipo 3) geram línguas L3.

Em outras palavras, o tipo de gramática induz uma ordem de inclusão de classes no conjunto das línguas possíveis. Por exemplo, a classe das línguas L3, gerada pelas gramáticas regulares, está contida na classe das línguas L2, gerada pelas gramáticas livres de contexto. Há, portanto, uma relação de inclusão entre as classes de línguas geradas por esses tipos de gramática. Cada uma dessas classes de línguas é mais restrita, menos geral, do que a classe que a antecede. Logo L3 $\subset$ L2 $\subset$ L1 $\subset$ L0 (o símbolo $\subset$ é lido como “está contida em”).

Gazdar & Mellish (*op. cit.*: 133) argumentam que, mesmo que existam línguas sensíveis ao contexto, a grande maioria das estruturas sintáticas atualizadas pelas línguas humanas pode ter como

105 Pesquisas têm apontado que a Gramática Léxico-Funcional é provavelmente deste tipo (*cf*. HARLOW & VINCENT, *op. cit.*: 10)

modelo uma gramática sintagmática livre de contexto.[106] Do ponto de vista lingüístico, encontra-se uma constatação plausível para a proposição dessa aproximação: Weinberg (1989), Partee *et al.* (*op. cit.*: 533) constatam que, desde a proposta da *Hierarquia de Chomsky*, a maioria dos pesquisadores tem proposto que o tipo de gramática das línguas naturais encontra-se entre os tipos 1 e 2, tendendo para o tipo 1, as gramáticas livres de contexto. Esse resultado é de extrema relevância para o estudo sobre o PLN, porque as técnicas de análise sintática criadas para as gramáticas livres de contexto revelam-se eficientes e elegantes quando aplicadas às línguas naturais. A figura abaixo representa as referidas relações de inclusão.



Para cada classe de **gramática** existe uma classe de **redes de transição** equivalente. Em outras palavras, dada uma gramática é

106 O dialeto do alemão falado em Zurique é citado por esses autores como um indício da existência de línguas sensíveis ao contexto.

possível construir uma rede de transição equivalente. Cada classe de rede de transição, por sua vez, pode ser interpretada como a descrição de uma classe de **autômatos**, isto é, máquinas computacionais abstratas que simulam os computadores. Nesse domínio, as redes de transição são denominadas diagramas de estado, isto é, diagramas que representam os estados internos de uma máquina computacional abstrata (GAZDAR & MELLISH, *op. cit.*). Assim, verificam-se as seguintes equivalências (SUDKAMP, *op. cit.*: 248):[107]

Tipos de língua	Tipos de gramática	Tipos de redes	Tipos de autômatos
0	Irrestrita	RTA	Máquina de Turing
1	Sensível ao Contexto	RTA	Autômato Limitado
2	GLC	RTR	Autômato de Pilha
3	Regular	RTEF	Autômato Finito

Analisadores gramaticais

Como assinalei anteriormente, o dispositivo computacional que executa a tarefa de análise gramatical em um SPLN é denominado *analisador gramatical*. Os *parsers* são, certamente, os dispositivos mais estudados no âmbito do PLN. Depois do trabalho pioneiro de Earley (*op. cit.*), merecem destaque os trabalhos de Berwick (1985), Dowty *et al.* (1985), Tomita (1986), Reyle & Rohrer (1987), Pritchett (1989), Pritchett & Reitano (*op. cit.*) e petrick (*op.cit.*).

O analisador gramatical, ao ser alimentado com uma seqüência de palavras, pode funcionar de duas maneiras: como um simples verificador ou como um verificador-construtor de estruturas

107 A RTR do exemplo acima, portanto, além de ser um formalismo equivalente à GLC exemplificada, é também a especificação de um autômato de pilha.

sintáticas. No primeiro caso, o dispositivo apenas sanciona a gramaticalidade da seqüência de palavras que recebe como entrada, assinalando se esta admite ou não uma estrutura sintática bem-formada.[108] No segundo caso, além de avaliar a gramaticalidade da seqüência de palavras submetida à análise, o analisador sintático constrói a representação sintática da frase ou, nos casos em que haja ambigüidades, as representações sintáticas admissíveis.

Quando implementados no computador, a gramática e o léxico têm a função de fornecer ao analisador gramatical informações sobre os objetos lingüísticos – itens lexicais, sintagmas, frases, relações sintáticas e semânticas, marcadores discursivos, entre outros – e as condições de boa-formação morfológica, sintática e semântica desses objetos. Assim, o analisador gramatical pode executar sua tarefa essencial: com precisão e sem ambigüidades, associar a cada objeto lingüístico sua categoria e estrutura correspondentes, de modo a explicitar suas relações de interdependência e reunir todas as informações que serão posteriormente processadas no nível pragmático-discursivo.

Esse processo automático de análise gramatical pode ser entendido como um tipo de resolução de problemas que envolve a exploração de um espaço de soluções possíveis. Essa tarefa pode ser especificada, levando-se em conta duas dimensões independentes: o sentido da análise e o modo de explorar hipóteses.

O sentido da análise pode ser (i) descendente (do inglês *top-down parsing*), caso em que o analisador, partindo do símbolo inicial

---

108 O analisador sintático pode ser considerado o grande responsável pela proliferação dos "asteriscos" que os lingüistas costumam colocar antes das estruturas sintáticas mal-formadas, em geral, empregadas para ilustrar as discussões teóricas.

da gramática, constrói agrupamentos sintáticos parciais até atingir os itens lexicais que compõem a seqüência de palavras, ou (ii) ascendente (do inglês *bottom-up parsing*), caso em que o analisador, partindo da seqüência de palavras, constrói agrupamentos sintáticos parciais até atingir o símbolo inicial da gramática.

Já o modo de explorar hipóteses pode ser: (i) exploração em profundidade (*depth-first search*), caso em que o analisador explora uma hipótese de cada vez, e (ii) exploração em paralelo (*breadth-first search*), caso em que o analisador explora mais de uma hipótese em paralelo.

A Gramática **G** abaixo, o Léxico **L** e a frase *Paulo viu Ana* ilustram esses conceitos.

Gramática G:

Regra 1:	F	→	SN SV
Regra 2:	SV	→	V SN
Regra 3:	SV	→	V

Léxico L:			
SN:	Paulo	V:	encontrou
	Pedro		matou
	Ana		viu
	a casa		viajou
	o carro		beijou

Passos	**ESTRATÉGIA 1** **Analisador gramatical descendente (*top-down parser*)** **Estrutura construída em profundidade (*depth-first*)**
1	(F(SN)(SV))
2	(F(SN(*Paulo*))(SV))
3	(F(SN(*Paulo*))(SV(V)(SN)))
4	(F(SN(*Paulo*))(SV(V(*viu*))(SN)))
5	(F(SN(*Paulo*))(SV(V(*viu*))(SN(*Ana*))))

Passos	**ESTRATÉGIA 2** **Analisador gramatical ascendente (*bottom-up parser*)** **Estrutura construída em profundidade (*depth-first*)**
1	(SN(*Paulo*))
2	(SN(*Paulo*))(V(*viu*))
3	(SN(*Paulo*))(SV(V(*viu*)))
4	(SN(*Paulo*))(SV(V(*viu*))(SN(*Ana*)))
5	(F(SN(*Paulo*))(SV(V(*viu*))(SN(*Ana*))))

Passos	**ESTRATÉGIA 3** **Analisador gramatical descendente (*top-down parser*)** **Estrutura construída em paralelo (*breadth-first*)**
1	(F(SN)(SV))
2	(F(SN(*Paulo*))(SV))
3	(F(SN(*Paulo*))(SV(V)(SN)))
4	(F(SN(*Paulo*))(SV(V(*viu*))(SN)))
5	(F(SN(*Paulo*))(SV(V(*viu*))(SN(*Ana*))))

Passos	**ESTRATÉGIA 4** **Analisador gramatical ascendente (*bottom-up parser*)** **Estrutura construída em paralelo (*breadth-first*)**
1	(SN(*Paulo*))
2	(SN(*Paulo*)) (V(*viu*))
3	(SN(*Paulo*)) (V(*viu*)) (SN(*Ana*))
4	(SN(*Paulo*)) (SV(V(*viu*))(Nome(*Ana*)))
5	(F(SN(*Paulo*))(SV(V(*viu*))(Nome(*Ana*))))

As estratégias assim apresentadas, entretanto, não deixam transparecer o complicado processo de exploração de hipóteses por que

passa um computador.[109] Para mostrar essa complexidade, exemplifico o processo com a ESTRATÉGIA 2.

Em primeiro lugar, o analisador lê a primeira palavra da seqüência *Paulo viu Ana* e compara com as palavras do léxico. Como há uma correspondência (*Paulo*=**Paulo**), o analisador substitui a palavra *Paulo* da seqüência que está analisando pelo símbolo categorial especificado no léxico, a saber SN. A seqüência passa, então, a ser **SN(*Paulo*) *viu Ana***. Para facilitar a exposição, todo símbolo categorial que for "transferido" para a frase será chamado "símbolo processado". Feita a substituição do item lexical pela categoria correspondente, o analisador recorre à gramática para verificar se todo o lado direito de alguma regra pode ser identificado com o símbolo processado. Como na gramática não há uma regra que preencha essa condição, o analisador é forçado a ler a próxima palavra. Depois de feita a identificação e a substituição, a frase passa a ter a seguinte forma: **SN(*Paulo*) V(viu) *Ana***.

Antes de continuar a descrição do processo, julgo oportuno chamar a atenção para o procedimento básico que foi executado pelo analisador. Observe que, na essência, o analisador repete o mesmo procedimento em dois momentos: identificação e substituição de símbolos. No primeiro momento, houve a identificação de itens lexicais, um pertencente ao léxico (**Paulo**) e o outro à frase (*Paulo*). Houve também uma substituição: no léxico, o símbolo que se encontra à direita (**Paulo**) foi substituído pelo símbolo que se encontra à esquerda (SN). No segundo momento, a identificação envolveu símbolos categoriais,

109 Na verdade, muitas das representações aqui apresentadas apenas fazem sentido quando implementadas em um computador. Por esse motivo, reconheço que esta tese, infelizmente, pode se tornar, em alguns momentos, também refém dos "formalismos bizarros" que denunciei nos capítulos iniciais, sobretudo, para um leitor não afeito a este universo.

que no caso não resultou em substituição, porque a condição de correspondência entre a categoria processada (SN) e o lado direito de regras da gramática não foi satisfeita. O importante é notar que essas duas operações básicas repetem-se ao longo de todo o processo de análise, ora identificando pares de itens lexicais, um pertencente à frase e o outro ao léxico, ora identificando um símbolo processado (ou uma seqüência de símbolos processados) com o lado direito de alguma regra da gramática. Como veremos, essas operações repetem-se no interior da própria gramática.

Retomando-se a descrição do processo, observa-se que o analisador, até este ponto da descrição, apresenta o seguinte resultado parcial: **SN(*Paulo*) V(*viu*) *Ana***. Em outras palavras, o analisador já "descobriu" que os dois primeiros elementos da frase pertencem às categorias SN e V, respectivamente. Neste ponto, antes de ler a próxima palavra, o analisador verifica, se há uma regra na gramática, cujo lado direito é do tipo SN V. Primeiro, porém, o analisador verifica novamente se há alguma regra cujo lado direito contenha apenas un SN. Como não há tal regra, a tentativa agora é encontrar uma regra, cujo lado direito contenha SN V. Como também não há regra desse tipo, a alternativa é, então, encontrar uma regra que contenha apenas V em seu lado direito. Com efeito, a Regra 3 preenche esse requisito. Assim, o analisador executa a substituição. A frase assume a seguinte forma: **SN(*Paulo*) SV(V(*viu*)) *Ana***.

Depois de verificar mais uma vez que o SN não pode ser substituído por nenhum outro símbolo e que o símbolo SV também está sujeito à mesma limitação, o analisador verifica se é possível agrupar os

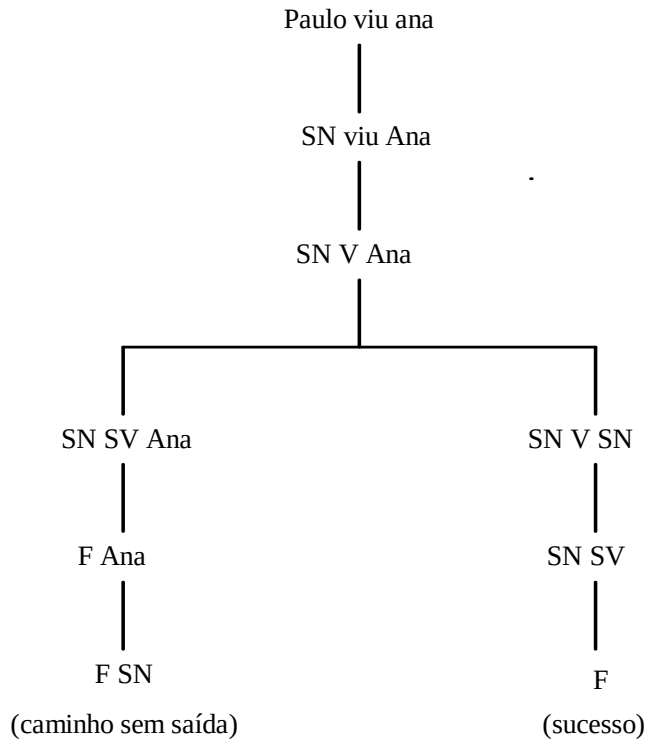
símbolos SN SV.[110] A Regra 1 autoriza essa substituição. Assim a frase passa à seguinte forma: **F(SN(*Paulo*)SV(V(*viu*))) *Ana***. Essa seqüência, por sua vez, leva o analisador a testar, uma a uma, as seguintes hipóteses: há alguma regra na gramática, cujo lado direito é da forma F? Há alguma regra na gramática, cujo lado direito é da forma F *Ana*? Há algum item lexical da forma *Ana*? Somente a última hipótese é confirmada. A frase passa então a ter a seguinte forma **F(SN(*Paulo*)SV(V(*viu*))) SN**. Exatamente neste ponto, o analisador não pode continuar o processo, porque o lado direito de nenhuma regra pode ser identificado com F, SN ou F SN. Se o mecanismo denominado "retrocesso" (*backtracking*) não for incorporado nesse analisador, ele simplesmente interrompe o processo e sinaliza que a frase não pôde ser reconhecida.

O mecanismo de retrocesso faz com que o analisador retorne ao ponto da análise em que havia outras alternativas que poderiam ter sido testadas. Em outras palavras, o analisador desfaz todas as operações até esse ponto e testa essas outras possibilidades. No exemplo, o analisador retrocede à situação em que a forma da frase era a seguinte: **SN(*Paulo*) SV(V(*viu*)) *Ana***. Neste ponto, o analisador tomou o "caminho errado" ao tentar agrupar SN SV. Essa decisão resultou num "caminho sem saída": F SN. Partindo, então, da forma **SN(*Paulo*) SV(V(*viu*)) *Ana***, o analisador novamente lê a palavra *Ana*, constrói a forma **SN(*Paulo*) SV(V(*viu*)) SN(*Ana)***, a partir da qual novas hipóteses serão levantadas e testadas. Depois de muitas outras tentativas e retrocessos mal sucedidos, o analisador monta a forma **SN(*Paulo*)**

110 Naturalmente, há a outra alternativa que é ler a palavra *Ana*. Essa alternativa, porém, não é testada, pois o analisador está programado para fazer todas as reduções possíveis antes de ler outros itens lexicais.

V(*viu*) SN(*Ana*) e, aplicando a Regra 2, chega à seguinte forma: **SN(*Paulo*) (SV(V(*viu*) SN(*Ana*))**. Finalmente, a Regra 1 autoriza a estrutura **F(SN(*Paulo*) (SV(V(*viu*) SN(*Ana*)))**, não sem antes verificar, mais uma vez, se o primeiro SN ocorre do lado direito de alguma regra da gramática.

Esse exemplo mostra, entre outras coisas, que o espaço de alternativas a ser explorado durante o processo de análise gramatical é, de fato, surpreendente. Logo, técnicas de exploração são o foco das pesquisas nesse campo. Pode-se representar parte desse espaço de alternativas por meio de uma árvore:



Os analisadores descendentes com mecanismos de retrocesso são comumente associados tanto às GLCs quanto às RTRs. Já os analisadores ascendentes são mais comumente associados somente às GLCs. Os analisadores mais eficientes, que empregam a estratégia

descendente e contêm uma estrutura de armazenagem de regras potenciais e estruturas sintáticas intermediárias já construídas – a "**estrutura de mapas do processamento**" (do inglês *chart structure*), além de poderem ser associados a esses formalismos, podem também ser associados às **gramáticas lógicas** (PEREIRA & WARREN, 1980; PEREIRA & SHIEBER, 1987; McCORD, 1980; ABRAMSON & DAHL, *op. cit.*). Esses analisadores, denominados **analisadores gramaticais munidos de mapas de processamento** (do inglês *chart parsers*), foram propostos por Kay (1985).

Esses procedimentos automáticos de análise gramatical que envolvem a exploração de todas as análises estruturais alternativas possíveis são denominados **não-determinísticos**, isto é, eles se encontram em situações em que é preciso tomar caminhos alternativos na busca de possíveis estuturas gramaticais bem-formadas. Como vimos, o analisador gramatical constrói a estrutura gramatical de uma frase passo a passo, fazendo hipóteses, eliminando hipóteses, retrocedendo, acumulando e descartando resultados parciais até atingir dois estados finais: sucesso ou falha. No primeiro estado, a frase é descrita como gramatical, situação em que o analisador apresenta como resultado as estruturas gramaticalmente especificadas pela gramática de que depende para funcionar. No segundo caso, depois de todas as tentativas, ele simplesmente sinaliza que não há nenhuma estrutura gramatical.

A discussão do determinismo no âmbito do PLN assume importância quando se considera a eficiência dos analisadores gramaticais. Como os analisadores não-determinísticos precisam enfrentar o problema da busca de soluções em um espaço de soluções alternativas, que pode se tornar extremamente grande em função das

inúmeras ambigüidades, tentar projetar **analisadores determinísticos** tem sido recentemente uma alternativa bastante promissora.

As pesquisas mostram que o falante, ao processar uma frase, não utiliza estratégias não-determinísticas. Em outras palavras, o processamento gramatical humano parece ser do tipo **determinístico**: ao invés de hipotetizar uma série de estruturas gramaticais intermediárias, ou até mesmo estruturas gramaticais completas e alternativas, e ir eliminando-as uma a uma até chegar àquela que corresponda à análise apropriada, o falante usa as informações de que dispõe e "automaticamente" (curiosamente!) determina **a** estrutura gramatical apropriada. O falante, durante o processamento de uma frase, parece "preferir" certas estruturas a outras, o que lhe permite descartar uma série de estruturas de imediato. Entre essas preferências há, por exemplo, a preferência por "escolher" um constituinte, cuja estrutura acrescente o menor número de nós possível à estrutura já processada (*minimal attachment*; FRAZIER & FODOR, *op. cit.*); a preferência por "acrescentar" o constituinte à estrutura já processada, de modo que ele fique mais à direita possível e no nível mais inferior possível (*right association*; FODOR & FRAZIER, *op. cit.*); e a preferência por acrescentar um constituinte seguindo as "preferências" de certos itens lexicais (*lexical preferences*; FORD, BRESNAN, & KAPLAN, 1982). Essas pesquisas procuram explicar por que certas interpretações de frases estruturalmente ambígüas, na ausência de qualquer outra informação prosódica, semântica e/ou pragmática, são preferidas a outras interpretações também possíveis.[111]

---

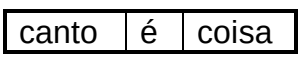
111 Resultados dessas pesquisas são também relevantes para evidenciar a importância do conhecimento sintático para o PLN, importância que nem sempre tem sido reconhecida. Hirst (*op. cit.*: 2) cita exemplos de pesquisadores que consideram a sintaxe um mero artefato inadequado para ser incorporado em um SPLN.

Acrescentando-se essas **técnicas determinísticas** aos analisadores gramaticais, é possível aumentar sua eficiência. Há dois tipos de analisadores determinísticos: os analisadores "*lookahead*" (MARCUS, *op. cit.*) e os analisadores "*shift-reduce*" (PEREIRA, 1985), ambos procuram eliminar as ambigüidades locais.

O analisador gramatical ascendente, proposto por Marcus, é composto de duas partes: regras gramaticais que especificam as ações do analisador e duas estruturas para manipular os dados. As regras gramaticais são regras de uma GLC. Uma das estruturas de dados é uma pilha (*pushdown stack*) que armazena estruturas sintáticas parciais temporariamente. A outra estrutura é uma memória temporária (*lookahead buffer*), com três posições, por onde as palavras da frase entram no analisador (da direita para a esquerda) e onde as estruturas que saem da pilha são colocadas. A primeira posição mais à esquerda corresponde à palavra que acabou de ser lida pelo analisador. As duas posições restantes correspondem às palavras a que o analisador tem acesso. Esse dispositivo (*lookahead*) permite que o analisador inspecione duas palavras adiante do ponto em que se encontra, antes de aplicar uma regra da gramática.

Essa informação é importante, porque restringe as possibilidades combinatórias que o analisador precisaria testar. Por exemplo, a ilustração abaixo mostra a situação do *buffer*, quando a palavra *canto*, na frase *Ele sabe que canto é coisa séria*, está sendo analisada:

b u f f e r



⇑

ponto em que o analisador se encontra

Note que, com essa informação adicional, o analisador é capaz de resolver a ambigüidade categorial. Como todo analisador sintático ascendente, antes de construir uma estrutura de nível superior na hierarquia sintática, esse analisador primeiro constrói as subestruturas. Por exemplo, antes de construir F (frase), ele constrói o SN e o SV. Esses resultados intermediários são armazenados na pilha na seguinte ordem: o SV sobre o SN. A operação de construção de F, também denominada redução, é feita no interior da pilha: o analisador simplesmente substitui os dois constituintes parciais pela estrutura oracional completa F. A ilustração abaixo ilustra o procedimento:

Antes da redução:

Pilha:

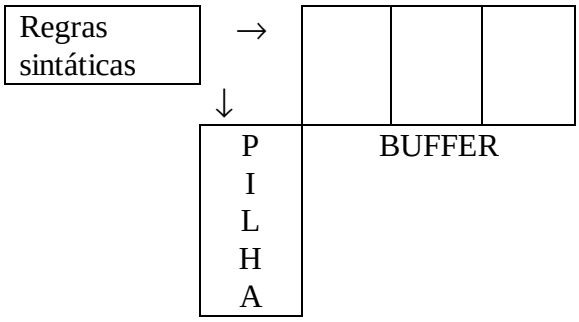
SV
SN

Depois da redução:

Pilha:

F = SN + SV

Esquematicamente o analisador pode ser assim visualizado:



Os analisadores *shift-reduce*, por sua vez, recorrendo a "oráculos" tornam-se determinísticos. Um oráculo é uma tabela, construída a partir da gramática, que especifica as ações e deslocamentos que o analisador deve executar. Observe o exemplo de KINDERMANN & MEIER, 1988: 136):

Regras da Gramática SR:

R1:	F	→	SN SV
R2:	SN	→	DET N
R3:	SN	→	SN SP
R4:	SP	→	P SP
R5:	SV	→	V
R6:	SV	→	V SN
R7:	SV	→	V SN SP

Oráculo construído a partir da Gramática SR (instruções que o analisador deve seguir)									
	Ações					**Ir para o estado x**			
Estados	**DET**	**N**	**P**	**V**	**$**	**F**	**SN**	**SV**	**SP**
0	sh3					**1**	**2**		
1					sucesso				
2			sh6	sh7				**5**	**4**
3		sh8							
4			r3	r3	r3				
5					r1				
6	sh3						**9**		
7	sh3				r5		**10**		
8			r2	r2	r2				
9			r4/sh 6	r4	r4				**4**
10			sh6		r6				**11**
11			r3	r3	r7/r3				

Os números à esquerda da tabela são os estados em que o analisador se encontra durante o processo de análise gramatical. Os símbolos **r** e **sh** são as abreviações de *reduce* e *shift*, respectivamente. O símbolo sh3 no estado 7, por exemplo, faz com que o analisador coloque o símbolo DET no *buffer*, que, neste caso, só contém uma única posição, e mude para o estado 3. O símbolo **$** assinala o final da frase. Os demais são os símbolos convencionais da gramática.

Subdomínio semântico

Uma vez caracterizadas as principais formas de representação das estruturas e processos morfossintáticos, passo para a questão da representação do significado, que, de acordo com a diretriz traçada no início deste capítulo, subdivide-se em duas partes: o problema da representação dos significados lexical e oracional, divorciada dos

contextos pragmático-discursivo e situacional, e o problema da ancoragem desses significados nesses contextos.

O subdomínio semântico focaliza a caracterização do significado da frase abstraída de seu contexto de ocorrência (cf. SCHUBERT & PELLETIER, 1982). Busca-se, nesse caso, a representação de uma espécie de forma lógica ou conteúdo proposicional da frase, tarefa que consiste em determinar o significado apropriado de cada item lexical em função do significado dos outros itens lexicais que integram a frase. Em outras palavras, derivar a forma lógica de uma frase é essencialmente caracterizar o valor semântico de cada item lexical, especificar as restrições de interpretação que cada um deles exerce sobre os outros e utilizar essas informações durante o processo de interpretação das frases.

Um primeiro modelo formal de semântica, que poderia ser cogitado, é a semântica de valor de verdade, desenvolvida por Montague (*cf*. DOWTY *et al.*, 1981). Esse modelo procura determinar as condições de verdade que precisam estar satisfeitas para que uma proposição seja verdadeira. Em outras palavras, essa semântica fornece meios para se especificarem os estados de mundo em que uma proposição é verdadeira. Os objetos semânticos dessa teoria são as entidades de um modelo matemático, isto é, indivíduos e relações que se estabelecem entre eles. Como as proposições engendradas pelas expressões lingüísticas não se restrigem ao mundo presente, Montague propõe os “mundos possíveis”, e chama de índice cada par “mundo possível-tempo”. Assumindo que o item lexical é a unidade básica do significado, estipula também que, em cada índice, haja uma correspondência que associe uma entidade do modelo a cada item lexical da língua.

Apesar do seu rigor formal, há que se ressaltar, entretanto, que esse modelo não é adequado para o PLN, porque a quantidade astronômica de elementos gerados por ele torna-o computacionalmente intratável. Em um universo composto de apenas duas entidades e dois índices, por exemplo, há $2^{2^{522}}$ elementos na classe das possíveis denotações das proposições e cada elemento é formado por um conjunto de 2^{512} de pares ordenados (*cf.* HIRST, *op. cit.*). Além disso, saber apenas as condições de verdade sobre uma determinada situação não é suficiente para especificar paráfrases e atitudes proposicionais e nem permite derivar conclusões com base naquilo em que acreditamos.

Embora não haja um modelo computacional-padrão para o tratamento semântico, HIRST (*op. cit.*: 26) propõe alguns parâmetros norteadores, argumentando que:

- a teoria subjacente deverá explicitar o papel da sintaxe no processo de construção do significado da frase, prevendo, por exemplo, um relacionamento entre a sintaxe e a semântica de tal forma que o analisador gramatical possa obter *feedback* durante o processamento sintático da frase;
- a teoria subjacente deverá também fornecer representações para os significados extensional e intensional, para os contextos opacos (em que a substituição de expressões referenciais são vetadas) e para o problema gerado pela referência genérica;
- a teoria subjacente deverá admitir algum tipo de composicionalidade;

- seus objetos semânticos deverão ser manipuláveis por regras de inferência e por procedimentos de resolução de problemas.[112]

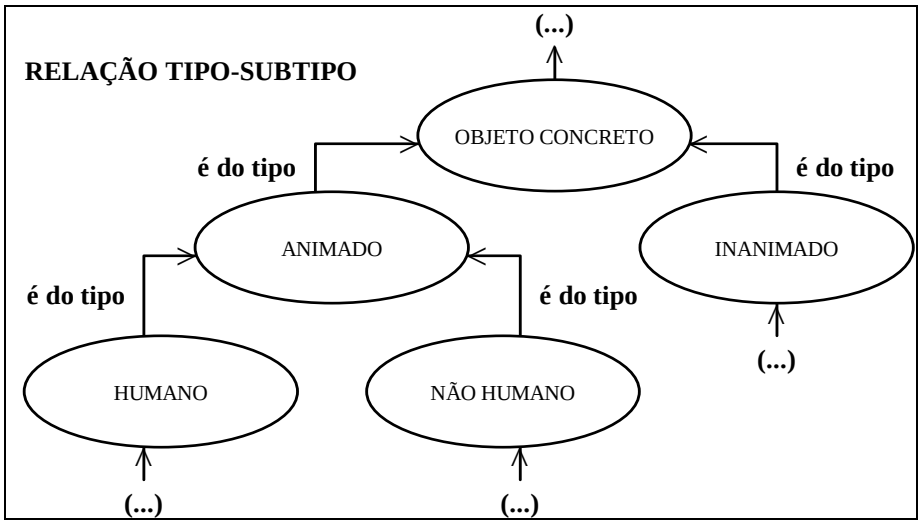
Diretrizes como essas têm motivado os pesquisadores a criarem linguagens artificiais, contendo a especificação das condições de verdade de uma proposição e regras de inferência que permitem deduzir proposições novas. Para esse fim, são propostas as **Redes Semânticas** (*cf.* QUILLIAN, *op. cit.*; BRACHMAN, 1979; WOODS, 1985).[113]

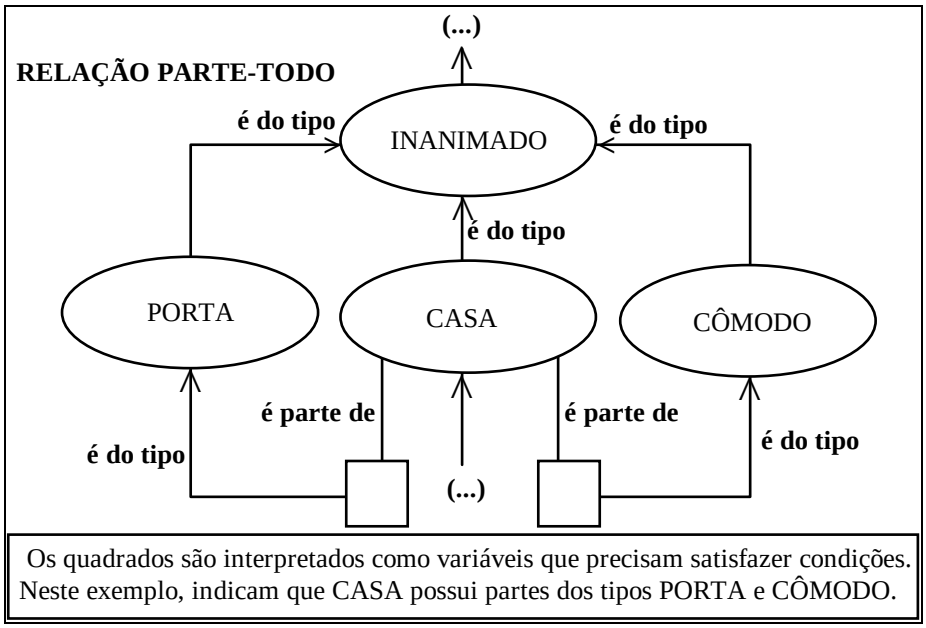
Uma rede semântica é basicamente um grafo composto de um conjunto de nós interligados por meio de arcos. Cada nó pode representar um **tipo** ou **subtipo** semântico ("os conceitos" de Jackendoff, por exemplo) e cada arco direcionado, que interliga pares de nós, pode representar tanto relações **tipo-subtipo** (as relações **é do tipo**) e **parte-todo** (as relações **é parte de**), como as **relações temáticas** (AGENTE, TEMA, etc.) que se estabelecem entre predicadores e seus argumentos. Os exemplos, a seguir, ilustram esses constructos:

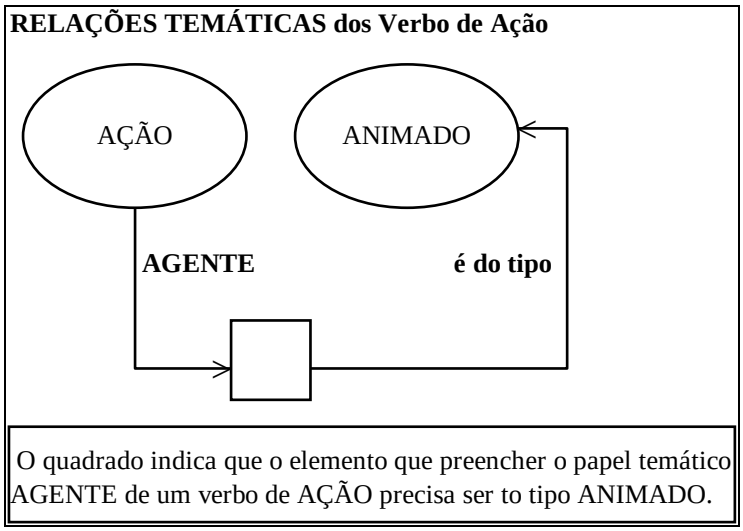
---

112 Os semanticistas Chierchia & Mcconnel-Ginet (*op. cit.*: 6-7), de fato, argumentam que parte da tarefa da semântica é especificar procedimentos algorítmicos de cálculo do significado da frase, a partir dos significados de itens lexicais e de constituintes oracionais

113 Quillian (*op.cit.*) propôs uma "memória semântica", em termos de redes, que contém entradas de um dicionário com as respectivas definições; um esquema de indexação associativa que permite ao sistema trilhar uma cadeia de referências indexadas. Dadas duas palavras, o sistema procura, se existir, uma associação de significados entre elas, exibindo a menor cadeia associativa que as une. Se, nas definições, há uma frase contendo as palavras A e B, e outra contendo B e C, a solicitação de relacionar A com C fará com que o sistema responda com as duas frases: uma delas contendo as palavras A e B, e a outra, as palavras B e C.


RELAÇÃO TIPO-SUBTIPO
(...)
OBJETO CONCRETO
é do tipo
é do tipo
ANIMADO
INANIMADO
é do tipo
é do tipo
(...)
HUMANO
NÃO HUMANO
(...)
(...)


RELAÇÃO PARTE-TODO
(...)
é do tipo
INANIMADO
é do tipo
é do tipo
PORTA
CASA
CÔMODO
é parte de
é parte de
é do tipo
é do tipo
(...)
Os quadrados são interpretados como variáveis que precisam satisfazer condições.
Neste exemplo, indicam que CASA possui partes dos tipos PORTA e CÔMODO.



Assim, as redes semânticas fornecem elementos para se representarem **conceitos** (os **tipos semânticos**) e **as restrições mútuas**, que os itens lexicais impõem uns aos outros, (os **papéis temáticos** e as **restrições selecionais**). Além disso, esses esquemas de respresentação abstratos possibilitam a codificação de múltiplas hierarquias de tipos semânticos, de restrições sobre o tipo semântico que deve preencher os papéis temáticos e de relações entre parte e todo. Desse modo, é torna-se possível construir tanto a representação do valor semântico das expressões lingüísticas em geral, quanto a representação da forma lógica de uma frase em particular.

Outro aspecto importante desse formalismo é que, além de permitir a representação de **tipos de estruturas** (*structure-types*), esse formalismo permite também representar **estruturas realizadas** (*structure-tokens*). Antes de apresentar, porém, um exemplo dessa possibilidade, considero oportuno recuperar alguns elementos do modelo de representação semântica proposto por Jackendoff.

Embora esse modelo, brevemente comentado no capítulo anterior, tenha sido proposto para explicar a estrutura do pensamento

humano, para os propósitos do PLN, essa adequação psicológica por ele imposta ao modelo torna-se absolutamente irrelevante. O importante, aqui, é destacar as distinções sugeridas entre **mundo projetado**, **conceitos**, **expressões lingüísticas**.

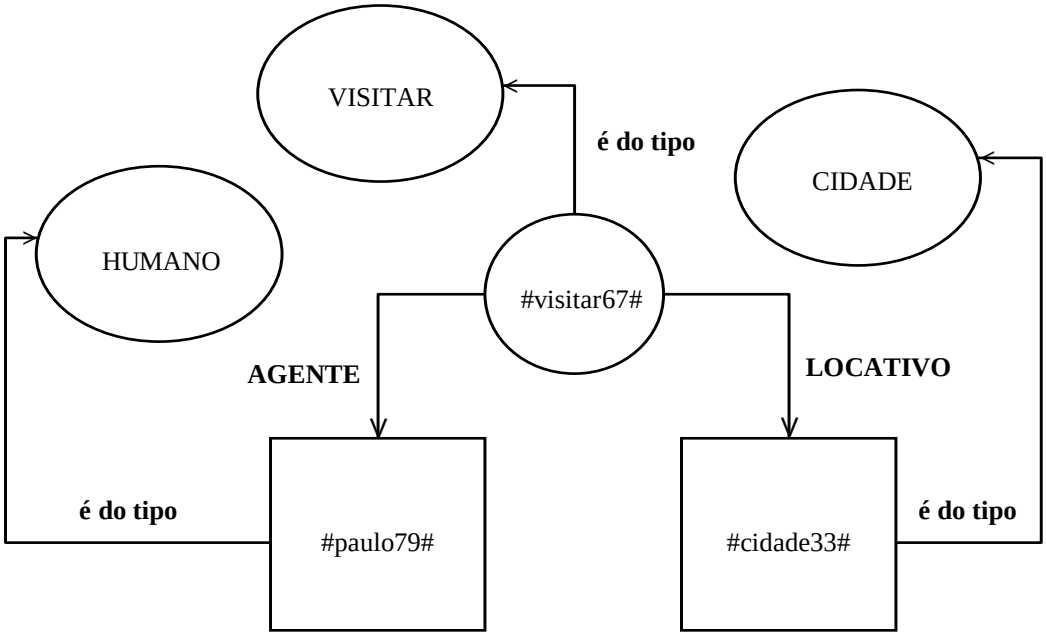
Se considerarmos que "o micromundo" a ser criado em um SPLN corresponde a uma espécie de mundo projetado e que a noção de tipo parece ser um constructo adequado para descrever conceitos, significados lexicais e relações semânticas, é possível postular que a expressão lingüística *casa* (**FORMA**) é do tipo CASA (**CONCEITO**) e que uma instância específica desse conceito como em, por exemplo, *Comprei uma casa*, é #casa17# (**REFERENTE**). Assim, especifica-se um modo de interpretar tanto o significado intensional como o significado extensional das formas da língua. Em outras palavras, é possível estabelecer dois tipos de correspondência: (i) entre a forma lingüística e o tipo semântico por ela expresso; (ii) entre o tipo semântico e um possível referente.

Em termos computacionais, estabelecer a correspondência (i) é associar à ocorrência *casa*, do exemplo acima, o tipo CASA. Já a correspondência (ii) é computacionalmente efetuada por meio de uma instrução que faz com que a máquina crie um símbolo específico para desinar o referente do sintagma nominal indefinido *uma casa*, que, neste exemplo, seria o símbolo #casa17#.

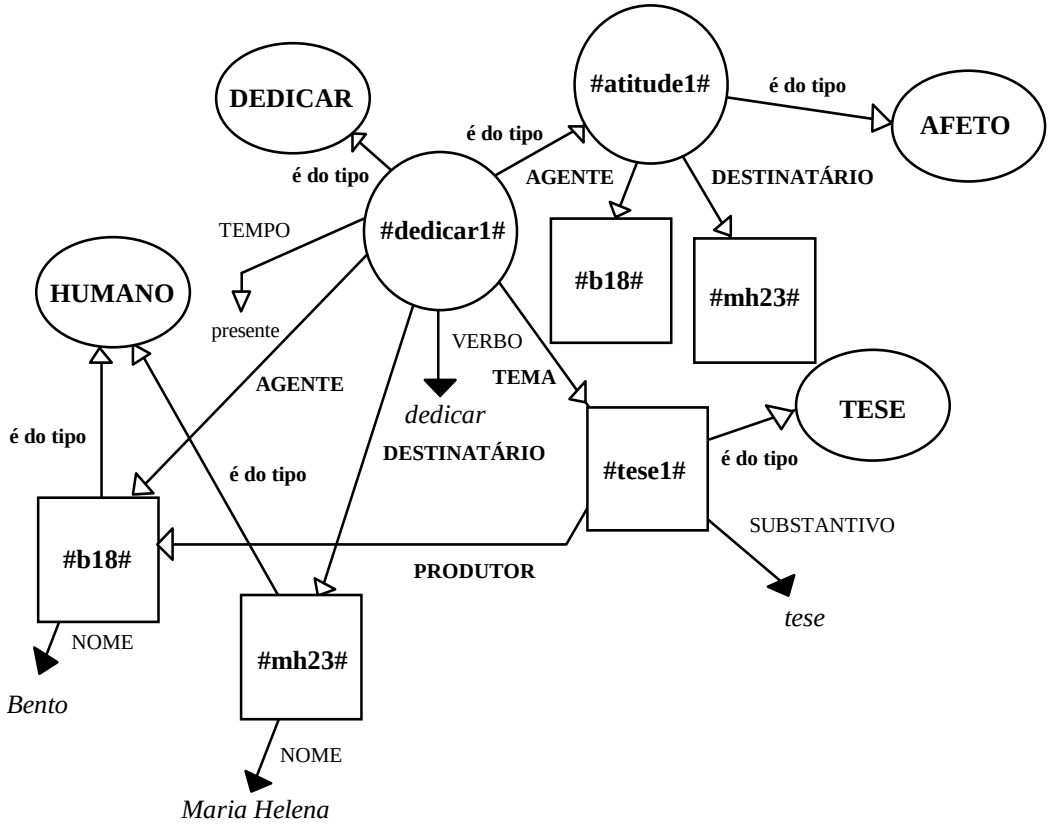
Vejamos outro exemplo. A partir da frase *Paulo visitou o Rio de Janeiro*, o sistema constrói uma forma lógica que, de modo simplificado, assume a seguinte representação VISITAR(PAULO,RIO DE JANEIRO). Essa representação é, posteriormente, associada a uma outra representação interna, que especifica os referentes e a ocorrência

específica do predicador: #visitar67#(#paulo79#,#cidade33#). Essa fórmula representa o evento específico e único #visitar67#, cujos participantes, também específicos, são #paulo79# e #cidade33#.

A rede semântica abaixo ilustra, então, o que acima denominei **estrutura realizada**:



A representação que utilizei como a dedicatória desta tese, repetida, a seguir, é uma rede semântica que representa o significado da frase *Bento dedica a tese à Maria Helena*.



Essa rede pode ser assim descrita: “trata-se de uma ocorrência específica #dedicar1# (lexicalizado pelo verbo *dedicar*) de um evento do tipo DEDICAR, que, então, “recebe por herança” três papéis temáticos: um AGENTE, preenchido por #b18# do tipo HUMANO (lexicalizado pelo NOME *Bento*), um TEMA, preenchido por #tese1# do tipo TESE (lexicalizado pelo substantivo *tese*) e um DESTINATÁRIO, preenchido por #mh23# do tipo HUMANO (lexicalizado pelo NOME *Maria Helena*). Como #dedicar1# é uma ocorrência de DEDICAR, #dedicar1# também “realiza por herança” #atitude1#, que, por sua vez, é do tipo AFETO e seus papéis temáticos, AGENTE e DESTINTÁRIO são preenchidos por #b18# e #m23#, respectivamente. Por fim, #tese1# recebe por herança do tipo TESE o papel temático PRODUTOR, preenchido por #b18#.”

Outro aspecto importante das redes semânticas é elas podem ser "traduzidas" para o cálculo de predicados. Observe as correspondências no quadro abaixo:

Redes Semânticas		**Cálculo de Predicados**
nós tipos	↔	predicados de um argumento
relações tipo-subtipo	↔	fórmulas
relações temáticas	↔	predicados de dois argumentos
valores dos papéis temáticos	↔	variáveis quantificadas pelo quantificador existencial ∃

Assim, a rede semântica que representa os verbos de ação pode ser expressa em termos das seguintes sentenças lógicas:

$$\forall x.\ \text{AÇÃO}(x) \Rightarrow \exists\ a.\ \text{AGENTE}(x,a)\ \&\ \text{ANIMADO}(a)$$

e

$$\forall a,x.\ \text{AÇÃO}(x)\ \&\ \text{AGENTE}(x,a) \Rightarrow \text{ANIMADO}(a)$$

Essa especificação lógica, por sua vez, permite formalizar os mecanismos de herança a que me referi no exemplo com o verbo *dedicar*. Como DEDICAR é um subtipo de AÇÃO, DEDICAR recebe, por "herança lógica", o papel temático AGENTE, associado ao tipo AÇÃO, e a restrição selecional ANIMADO, associada ao papel temático.

O estabelecimento de hierarquias dessa natureza representa uma grande economia no processo de descrição dos elementos, porque as características gerais de um determinado tipo precisam ser codificadas uma única vez, dado que o mecanismo de herança se encarrega de transferir essas características gerais para os subtipos. Graças a esse mecanismo, basta especificar que DEDICAR é um tipo de AÇÃO. As

regras de inferência propostas acima fornecem as informações complementares. A fórmula abaixo ilustra a semântica de dedicar:

$$\forall x.\ \mathbf{DEDICAR(x) \Rightarrow AÇÃO(x)}\ .$$

Há que se ressaltar, contudo, que esse tipo de dedução lógica, embora seja suficiente para formalizar vários aspectos e fenômenos das línguas naturais, revela-se inadequado para formalizar o que se denomina **inferência por omissão** (do inglês *default reasoning*), ou "inferência plausível" (PERLIS, 1990; CARPENTER & THOMASON, *op. cit.*). Assim, ao lado de inferências lógicas, um SPLN precisa simular também inferências que podem gerar conclusões conflitantes. Esse tipo de inferência, que geralmente toma por base o senso comum, pode ser ilustrado com um exemplo.

Suponha, primeiro, que um SPLN possua a seguinte **regra de inferência por omisssão**:

[a] $\forall x.\ AVE(x) \rightarrow^{o} VOAR(x)$

Essa regra dever ser assim interpretada:

- se "x é uma ave" é uma proposição **verdadeira** e **não** se pode provar que "x voa" é uma proposição **falsa**, então, concluir que "x voa" é uma proposição **verdadeira**.

Agora, suponha que o sistema receba as seguintes informações:

[b] $\forall x.\ PINGÜIM(x) \rightarrow \neg\ VOAR(x)$

[c] AVE(#tictac#)

Essas informações dizem ao sistema que “pingüins não voam” e que “tictac é uma ave”. A partir dessas informações, o sistema pode concluir que:

[d] VOAR(#tictac#)

Finalmente, suponha que uma nova informação sobre a ave tictac seja:

[e] PINGÜIM(#tictac#)

Neste ponto, o sistema precisa tomar uma decisão, uma vez que há duas informações verdadeiras ([c] e [e]) e as duas regras [a] e [b] são perfeitamente aplicáveis, resultando, porém, em duas proposições contraditórias. Aplicando-se a regra [a], obtém-se: VOAR(#tictac#); aplicando-se a regra [b], ¬ VOAR(#tictac#).

Como a regra [a] não é uma regra da lógica clássica, mas da **lógica por omissão** (do inglês *default logic*), o conflito é resolvido, levando-se em conta que, diante da informação específica, isto é, PINGÜIM(tictac), a regra [a] pode ser ignorada. Assim, o sistema registra AVE(#tictac#) e ¬ VOAR(#tictac#).

Como a “linguagem dos *frames*” é equivalente à linguagem lógica (*cf.* HAYES, 1979), uma outra possibilidade é representar as redes semânticas em termos desses constructos. Um *frame* é simplesmente uma estrutura formada por **atributos**, **valores** e **restrições** sobre os

elementos que podem se tornar valores dos atributos. Uma característica importante dos *frames* é que cada um de seus atributos é uma **função** que toma o objeto descrito pelo *frame* e produz o valor apropriado. Da mesma forma que existem redes semânticas tipo e redes semânticas ocorrência, distinguem-se também *frames genéricos* e *frames realizados*, espelhando-se assim a distinção tipo-ocorrência, ou tipo-realização.

O *frame* genérico **frame**-1, a seguir, representa o tipo CORPO HUMANO:

(frame-1 +CORPO HUMANO (X)
(**cab** +CABEÇA)
(**tro** +TRONCO)
(**mem**+MEMBROS))

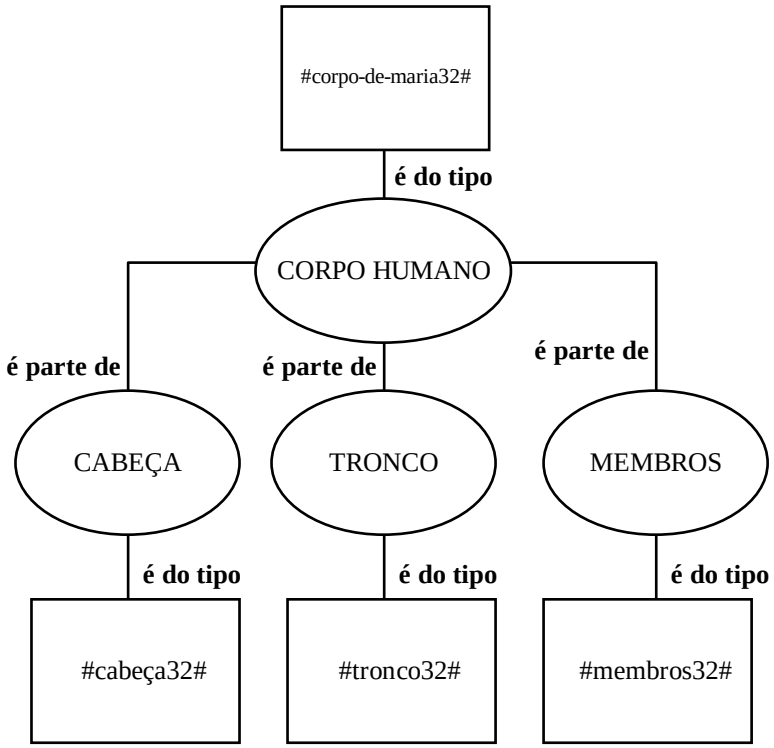
Essa estrutura representa que todos os objetos do tipo CORPO HUMANO possuem artributos do tipo CABEÇA, TRONCO e MEMBROS, identificáveis por meio das funções **cab**, **tron**, e **mem**, respectivamente. Já o *frame* abaixo representa uma possível realização do **frame**-1. Em outras palavras, aplicando-se **frame**-1 ao valor #corpo-de-maria32#, isto é, **frame**-1(#corpo-de-maria32#), obtém-se:

(**frame-1** CORPO HUMANO (#corpo-de-maria32#)
(**cab** #cabeça32#)
(**tro** #tronco32#)
(**mem** #membros32#))

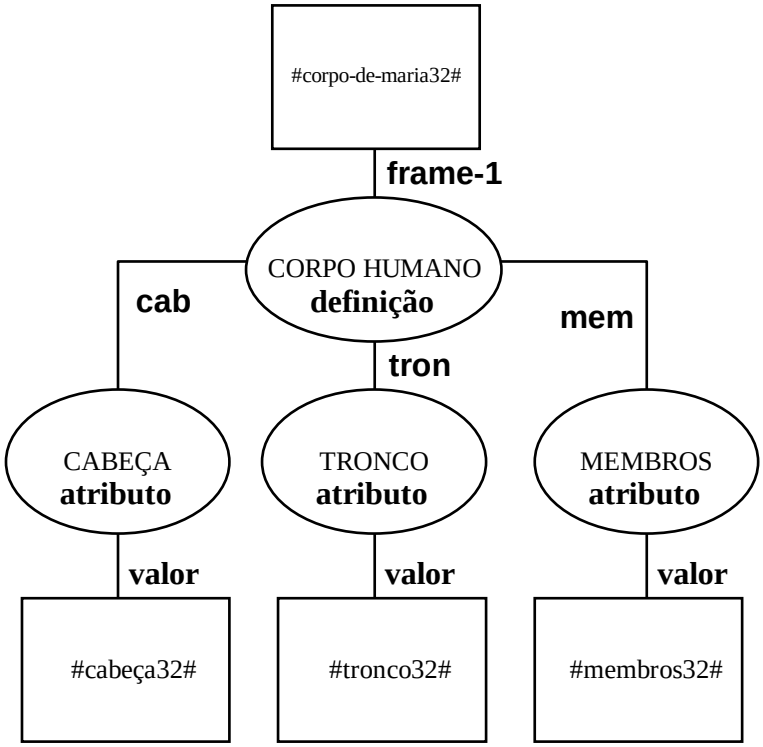
O valor da função **cab**, por exemplo, quando aplicada ao valor #corpo-de-maria32#, isto é, **cab**(#corpo-de-maria32#), resulta em #cabeça32#.

Observe as equivalências entre redes semânticas e *frames*:

Rede Semântica:



Frame: (frame-1 (#corpo-de-maria32#)



Para finalizar, apresento duas representações em termos de *frames*, tomando por base a proposta de NIRENBURG *et al.* (*op. cit.*). O primeiro *frame* representa a entrada lexical para o adjetivo *ruim* e o segundo, uma regra sintagmática.

```
(ruim (frame   +ruim-adj
                  (CATEGORIA (valor adj))
                  (SINTAXE
                        (atributivo + ))
                  (ESTR-SINTÁTICA
                        (       (raiz $var1)
                                (categoria n)
                                (mods ((raiz $var0)))))
                  (SEMÂNTICA  (PROJEÇÃO-LEXICAL
                                (^$var1
                                (instância-de (semântica (*OR*
                                                *digerível
                                                *atributo-olfativo
                                                *atributo-gustativo
                                                *atributo-visual))))
                        (ATITUDE
                                (tipo (valor disfórico))
                                (valor-da-atitude (valor 0.8))
                                (escopo (valor ^$var1))
                                (atribuída-a (valor *produtor*)))))))
```

```
(<DECLARATIVA> <==> (<SN> <SV>)
        (       ((x1 caso) = nom)
                ((x2 forma) =c finita)
                (*OR*
                        (       ((x2) :tempo) = presente)
                                ((x1 conc) = (x2 conc)))
                        (       ((x2 :tempo) = passado)))
                (x0 = x2)
                ((x0 suj) = x1
                ((x0 :modo) = dec))))
```

Subdomínio pragmático-discursivo

Além da representação do significado, é preciso também reconhecer a necessidade de ancoragem desse significado nos contextos pragmático-discursivo e situacional. Trata-se de reconhecer o problema da contextualização da forma lógica, que exige, além da representação do discurso e de sua manifestação em textos, a representação de seus participantes, com suas "visões de mundo", e a especificação de conceitos que possibilitem, pelo menos, a representação de parcelas de "conhecimento de mundo".

O conhecimento de mundo desempenha papel fundamental no processo de interpretação dos textos. Problemas básicos, como o fenômeno da ambigüidade e o estabelecimento dos referentes de expressões referenciais, só podem ser solucionados, ou pelo menos encaminhados, com o auxílio desse tipo de conhecimento. Embora haja diversas formas de "representação do conhecimento", dependendo dos tipos de predicados e de inferências utilizados, todas elas fazem a distinção entre proposições e termos. Aquelas referem-se a tudo o que pode ser considerado verdadeiro ou falso, e estes, a tudo o que pode

representar quaisquer tipos de objetos (objetos concretos, eventos, espaço, tempo e idéias, por exemplo).

Os sistemas de representação do "conhecimento", no sentido técnico aqui empregado, além de incluir as representações do conhecimento lingüístico – as representações dos conhecimentos lexical, morfológico, sintático, semântico, pragmático e discursivo, canonicamente estudados no âmbito da teoria lingüística, inclui também as representações de conhecimentos extralingüísticos.[114] Estas referem-se a complexos de representações que incluem representações dos agentes do discurso, de suas crenças, de coordenadas espácio-temporal, de parcelas do conhecimento de mundo que os agentes possuem, que, por sua vez, incluem o conhecimento de situações gerais e específicas, bem como o conhecimento de ações físicas e abstratas.

É desnecessário dizer que o conhecimento sobre as atividades do cotidiano é essencial para a compreensão de textos narrativos. A resolução de ambigüidades lexicais e estruturais e a identificação de referentes das expressões referenciais são processos que dependem desse conhecimento. Além disso, esse conhecimento é também essencial para responder a perguntas que demonstram um nível básico de compreensão de narrativas. As técnicas empregadas para a representação desse tipo de informação contextual variam desde a aplicação de estruturas rígidas, como os *scripts*, até os sistemas que utilizam inferências no processo de explicitação do atos de fala, sistemas que recorrem à informação sobre ações, planos e objetivos ("os sistemas de planejamento de ações").

---

114 As duas formas de "representação do conhecimento" mais utilizadas são a lógica e os *frames* (*cf*. BRACHMAN & LEVESQUE, *op. cit.*).

No subdomínio pragmático-discursivo, portanto, entra em jogo um complexo de representações (BRACHMAN & LEVESQUE, *op. cit.*; VELDE, *op. cit.*; VIEHWEGER, 1989; CONTE, 1989; KAYSER, 1989; SCHA, *op. cit.*) – como os *scripts* (SCHANK & ABELSON, *op. cit.*; LEHNERT, 1979), *frames* (MINSKY, 1975; SIDNER, 1979; CHARNIAK, 1979; HAYES, 1979), *plans* (COHEN & PERRAULT, *op. cit.*; PERRAULT & ALLEN, 1980; APPELT, *op. cit.*, McKEOWN, *op. cit.*), representação da estrutura do discurso (GROSZ & SIDNER, *op. cit.*; WEBBER, 1987 e 1990), do foco de atenção dos participantes do discurso (GROSZ, 1977; SIDNER, 1983), organização da memória (SCHANK, *op. cit.*).

Para evitar detalhamentos infindáveis, restrinjo a discussão a apenas algumas propostas que exemplificam aspectos essenciais para a compreensão de questões deste subdomínio, que aborda um tema que tem sido o grande desafio para a teoria lingüística e preocupação primeira e constante para o PLN.

Inspirados nas estratégias de planejamento (do inglês *planning*), desenvolvidas pelos estudos de Inteligência Artificial (voltados a manipulação de condutas dos robôs), os pesquisadores do PLN, reconhecendo – concordando com os filósofos da linguagem – que "linguagem é também ação", procuraram adaptar este modelo para a representação dos atos de fala. Foram então desenvolvidos *planners* ("sistemas de planejamento de ações") também para o discurso – sistemas computacionais que visam a detalhar as intenções subjacentes aos enunciados – que trouxeram um avanço considerável para o equacionamento deste subdomínio representacional.

No contexto do PLN, parte-se dos seguintes pressupostos: [i] os agentes comunicativos, o computador e o usuário pretendem alcançar certos objetivos; [ii] para alcançá-los, esses agentes constroem um plano de ação em termos de um ou mais atos de fala; [iii] o interlocutor, no processo de identificação dos atos de fala, é capaz de reconstruir, pelo menos, parte dos objetivos pretendidos pelo locutor.[115]

Allen & Perrault (*op. cit.*) propõem modelos que simulam o processo de identificação dos atos de fala. As crenças são representadas por meio de redes semânticas. As crenças de cada agente comunicativo induzem, sobre essas redes, partições, denominadas espaços de crenças. Reconhecem-se dois tipos de crenças: explícitas, proposições em que os agentes efetivamente acreditam, e implícitas, proposições em que os agentes poderiam acreditar.

Esses problemas têm sido parcialmente solucionados, empregando-se a metodologia desenvolvida nos sistemas de planejamento e nas gramáticas de unificação funcional, respectivamente. Cohen & Perrault (*op. cit.*) mostram como usar os atos de fala em um sistema de planejamento do discurso, propondo um modelo abstrato em que um determinado objetivo comunicativo, previamente previsto para o sistema (fornecer uma explicação sobre algum tópico, por exemplo), é alcançado por meio de uma seqüência de atos de fala. Appelt (*op. cit.*) elabora um outro modelo que especifica como as seqüências de atos de fala são convertidas em frases. McKeown (*op. cit.*), empregando representações em termos de *scripts*, propõe um modelo de discurso que permite ao sistema gerar textos com extensão de um parágrafo. Para isso,

---

115 O modelo computacional para os atos de fala foi proposto inicialmente por Cohen & Perrault (*op. cit.*). Allen & Perrault (*op. cit.*) empregaram o mesmo modelo em sistemas de "perguntas e respostas". A transposição desse modelo para um modelo computacional da estrutura do discurso encontra-se nos trabalhos de Grosz & Sidner (*op. cit.*).

desenvolve um *planner* que caracteriza três estruturas, ou esquemas textuais, que correspondem aos três objetivos comunicativos: definir, comparar e descrever. Esses três esquemas são empregados para nortear o processo de geração de texto pelo computador. A implementação dos esquemas foi feita empregando uma RTA. O sistema implementado é capaz de gerar parágrafos como respostas a consultas sobre informações contidas em um banco de textos.

O tratamento das expressões lingüísticas referenciais é outro tema que merece destaque. O dispositivo computacional desenvolvido para esse fim recorre à elaboração de listas que registram todos os referentes mencionados em segmentos do discurso já processados. Esses referentes podem ser utilizados no processo de interpretação dos elementos anafóricos ou elididos. Como vimos no capítulo anterior, Grosz & Sidner (*op. cit.*), que dividem o problema da representação do discurso em três subproblemas – segmentação, foco de atenção e intenção dos participantes do discurso –, procuram determinar propriedades do discurso que desempenham papel decisivo na determinação de referentes dos vários tipos de sintagmas nominais definidos. Essas autoras mostram a estreita ligação entre o processo de focalização e a determinação de referentes das expressões lingüísticas.

Charniak (1973) mostra a necessidade de incluir, no processo de determinação dos referentes das expressões referencias, informação do senso comum e cadeias de inferências dela decorrentes. Argumenta que por mais sofisticadas que sejam, as restrições puramente sintáticas e semânticas não são suficientes para especificar os referentes univocamente.

O modelo formal que Nirenburg *et al*. (*op. cit.*) propõem, para representar o **Significado do Texto** (ST), serve para ilustrar uma tentativa de integração dos três domínios: gramatical, pragmático e discursivo. O significado do texto é representado por meio de quatro sub-estruturas: **Conteúdo Proposicional** (P), **Relações** entre elementos intra e interfrasais e entre segmentos do discurso (R), **Atitudes Proposicionais** dos agentes comunicativos, bem como dos agentes representados no discurso (A), e **Intenções** do produtor do texto (I). As relações são classificadas segundo a seguinte tipologia:

- relações entre os elementos do domínio do discurso (relações de dependência entre eventos, estados e objetos; relações conjuntivas entre elementos adjacentes, relações de escolha, relações de co-referência entre elementos textuais, relações temporais);
- relações entre elementos do texto (parte-todo, paráfrase, ou reformulação, e conclusão);
- relações entre intenções e os componentes do texto relacionados ao domínio do discurso (tempo da fala e tempo do evento)

A representação das sub-estruturas ST = {P,R,A,I} são resumidas, a seguir, representando o significado abstrato do "mini-texto": *Pedro viu Maria e correu*.

P={P1,P2,...Pn} [conteúdo proposicional]

P1= {conceito instanciado (c1),papel temático associado a um conceito (θ1),aspecto (a1), tempo(t1)}

P1={EVENTO PERCEPTIVO, EXPERIENCIADOR = PEDRO, TEMA = MARIA , PERF, PAS}

P2={EVENTO CINÉTICO, AGENTE=PEDRO,PERF,PAS}

R={R1,R2,...,Rn}[relações entre elementos intra e interfrasais e entre segmentos do discurso]

R1={relação tipo1, argumentos1, valor da relação1}

R1:{DOMÍNIO-CAUSA, EVENTO PERCEPTIVO → EVENTO CINÉTICO,0}
R2:{DOMÍNIO-CO-REFERÊNCIA, PEDRO/pro,0}
R3:{DOMÍNIO-TEMPORAL-DEPOIS,TEMPO2/TEMPO3,0}

A={A1,A2,...,An}[atitudes proposicionais]

A1={tipo1,valor1,atribuída-a1,escopo1,tempo1}

A1 = {MEDO, 0.9,ATRIBUÍDA-A PEDRO,ESCOPO: EVENTO PERCEPTIVO, TEMPO}
A2 = {CRENÇA,1,ATRIBUÍDA-A PRODUTOR,ESCOPO: R1}

I={I1,I2,...I3}[intenção do produtor do texto]

I1={conceito que especifica um ato de fala1,escopo1)

I1={ATO-INFORMATIVO, ESCOPO: R1}.

A título de conclusão deste capítulo, apresento um esquema geral que contém a síntese de temas que merecem ser analisados do ponto de vista do PLN:

	TEMAS	
MORFOLOGIA		**SINTAXE**
Representação das formas primitivas Formas flexionais Formas derivacionais Formas clíticas		Representação das estruturas sintáticas bem formadas Categorias sintáticas Constituintes oracionais Tipos oracionais Expedientes sintáticos
	Representação do Léxico	
SEMÂNTICA		**PRAGMÁTICA e DISCURSO**
Representação do significado lexical e proposicional Semântica lexical Semântica frasal		Representação do significado em contexto Anáfora Elipse Dêixis Estrutura informacional Atos de fala Causa-efeito Tempo e aspecto

CAPÍTULO 6 – Equacionamento do domínio implementacional

> *"Primary issues in natural language processing include characterizing the capabilities that various components of a language processing system should have, what form they should take, what part they play in processing, and how they should be organized to ensure that they play their parts effectively."*
>
> B. Grosz, K. Jones & B. Webber (1986: xii)

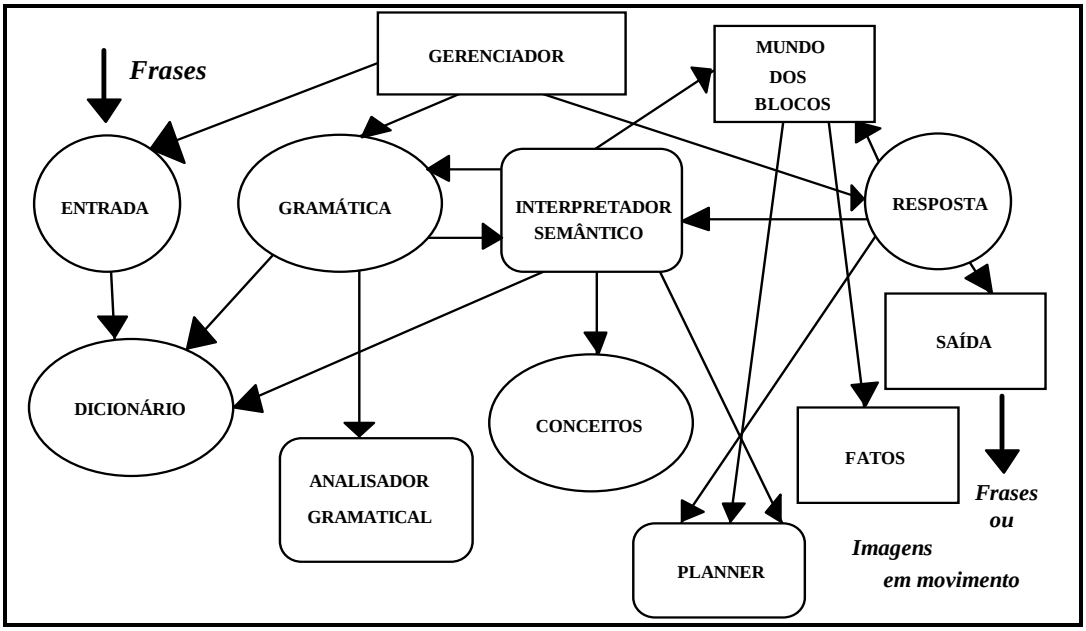
A discussão das questões do domínio implementacional ficarão restritas a três aspectos de seu equacionamento que considero relevantes para o programa de pesquisa integrado do PLN que venho delineando neste trabalho:

- a organização de um SPLN típico;
- o fluxo de informações que circulam no seu interior;
- a caracterização de um "ambiente computacional" para o seu desenvolvimento.

Como discutir esses aspectos implica delinear o equacionamento global de um SPLN, este capítulo torna-se oportuno para apresentar os seus vários componentes e permitir ao leitor uma visão de conjunto do empreendimento. Acredito que, para isso, o programa "mundo dos blocos" de Winograd é bastante ilustrativo para iniciar as discussões.

O “mundo dos blocos” de Winograd

Winograd projetou um SPLN com doze módulos, isto é, doze macroprogramas de computador escritos na linguagem de programação LISP: GERENCIADOR, ENTRADA, GRAMÁTICA, INTERPRETADOR SEMÂNTICO, RESPOSTA, ANALISADOR GRAMATICAL, DICIONÁRIO, CONCEITOS, MUNDO DOS BLOCOS, SAÍDA, PLANNER e FATOS, inter-relacionados conforme o diagrama a seguir.[116]



O macroprograma GERENCIADOR simplesmente ativa os macroprogramas ENTRADA, GRAMÁTICA e RESPOSTA.

O macroprograma ENTRADA recebe as frases digitadas pelo usuário, consulta o DICIONÁRIO, executa a análise morfológica, seleciona as estruturas de traços sintáticos e as “definições semânticas”

---

116 Winograd empregou as seguintes denominações: MONITOR, INPUT, GRAMMAR, SEMANTICS, ANSWER, PROGRAMMAR, DICTIONARY, SEMANTIC FEATURES, BLOCKS, MOVER, PLANNER e DATA, respectivamente.

de cada palavra que compõe a frase e, por fim, transfere os resultados para o módulo GRAMÁTICA.

O macroprograma GRAMÁTICA é o módulo coordenador principal de todo o processo de interpretação das frases e está ligado aos módulos DICIONÁRIO, ANALISADOR GRAMATICAL e INTERPRETADOR SEMÂNTICO. Este último fornece a interpretação semântica da frase com o auxílio dos módulos GRAMÁTICA, DICIONÁRIO, CONCEITOS, MUNDO DOS BLOCOS e PLANNER.

O macroprograma RESPOSTA controla as respostas fornecidas pelo sistema e mantém um registro do discurso para eventuais referências futuras com a ajuda dos módulos INTERPRETADOR SEMÂNTICO, MUNDO DOS BLOCOS, SAÍDA e PLANNER.

O macroprograma ANALISADOR GRAMATICAL é o módulo responsável pela execução da análise sintática e pela sua representação gráfica.

O macroprograma DICIONÁRIO é um módulo composto de duas partes: a primeira contém os traços sintáticos associados a cada item lexical; a segunda contém uma representação semântica para cada item.

O macroprograma CONCEITOS fornece a estrutura conceitual abstrata; por meio dela os itens e expressões lexicais são interpretados.

O módulo MUNDO DOS BLOCOS é um macroprograma que contém o conhecimento do sistema sobre as propriedades do mundo físico em que opera. Esse componente também ‘‘sabe’’ como alcançar objetivos em seu mundo fechado e como deduzir fatos novos a partir de fatos conhecidos.

O macroprograma SAÍDA é o módulo responsável pela movimentação dos elementos gráficos do "mundo dos blocos'' (o braço do robô e as figuras geométricas tridimensionais) e pela impressão, no monitor, tanto das frases digitadas pelo usuário como das respostas geradas pelo sistema.

O módulo FATOS representa a ''memória visual'' do sistema. Nele, os fatos sobre a cena são registrados: os objetos, seus tamanhos, formas, cores e localizações.

Por fim, o macroprograma PLANNER, o ''cérebro'' dedutivo do sistema, é o responsável pelo controle do processamento, orientando o módulo ANALISADOR GRAMATICAL e deduzindo novos fatos que alimentarão o módulo MUNDO DOS BLOCOS.

Uma arquitetura para um SPLN

Conceitualmente, as arquiteturas propostas para os sistemas de PLN acabam por espelhar a arquitetura proposta para o sistema lingüístico humano (*cf*. FRAZIER, 1989: 26). Como um sistema lingüístico, um SPLN deve possuir módulos autônomos que realizem tarefas específicas e especializadas, e módulos que armazenem um modelo de conhecimento proposicional que visa a criar simulacros de parcelas de mundo que lhe servem de referencial para interpretar os enunciados lingüísticos. Como os falantes de uma língua, o SPLAN, além de fazer inferências lógicas e, portanto, precisas, necessita também fazer inferências plausíveis, baseadas no senso comum (BONISSONE, 1990; DAVIS, 1990; NUTTER, 1990).

Máquinas inteligentes ?

Os pesquisadores da inteligência artificial costumam dizer que a maneira mais estruturada de se "transferir inteligência" para os computadores é criar programas modulares. Nesses programas, cada módulo contribui acumulativamente para a "inteligência" global do sistema.[117] Cada módulo, individualmente, é um pouco "menos inteligente" que o programa todo. Além disso, cada módulo é também subdividido em fragmentos menores, "menos inteligentes" que o próprio módulo. Finalmente, há partes menores ainda, totalmente desprovidas de qualquer vestígio de inteligência. Elas simplesmente obedecem mecanicamente às regras que devem executar. Em outras palavras, cada parcela do programa é um "pequeno especialista", limitado à resolução de problemas compatíveis com a quantidade de "inteligência" com que foi alimentado. A partir dessas considerações, é possível se estabelecer uma métrica para o grau de sofisticação dos SPLNs.

Como a interpretação dos enunciados lingüísticos necessariamente exige o domínio de conhecimentos extralingüísticos, as representações e o modo de utilização desses conhecimentos, além de serem essenciais para o desempenho adequado dos SPLNs, servem também de critério para classificá-los:

- sistemas que usam estruturas *ad hoc* sem representação de qualquer tipo de conhecimento prévio: a informação a ser processada é traduzida para alguma representação lógica interna que

117 Essa "inteligência", concebida como a capacidade de resolver problemas, é medida em função da quantidade e da qualidade das informações e dos mecanismos que o sistema dispõe para realizar suas tarefas. Quanto mais representações do mundo físico e conceitual e quanto mais regras de inferência e estratégias de manipulação dessas representações o sistema possuir, mais equipado estará para realizar as mais complexas tarefas que dependam da linguagem humana.

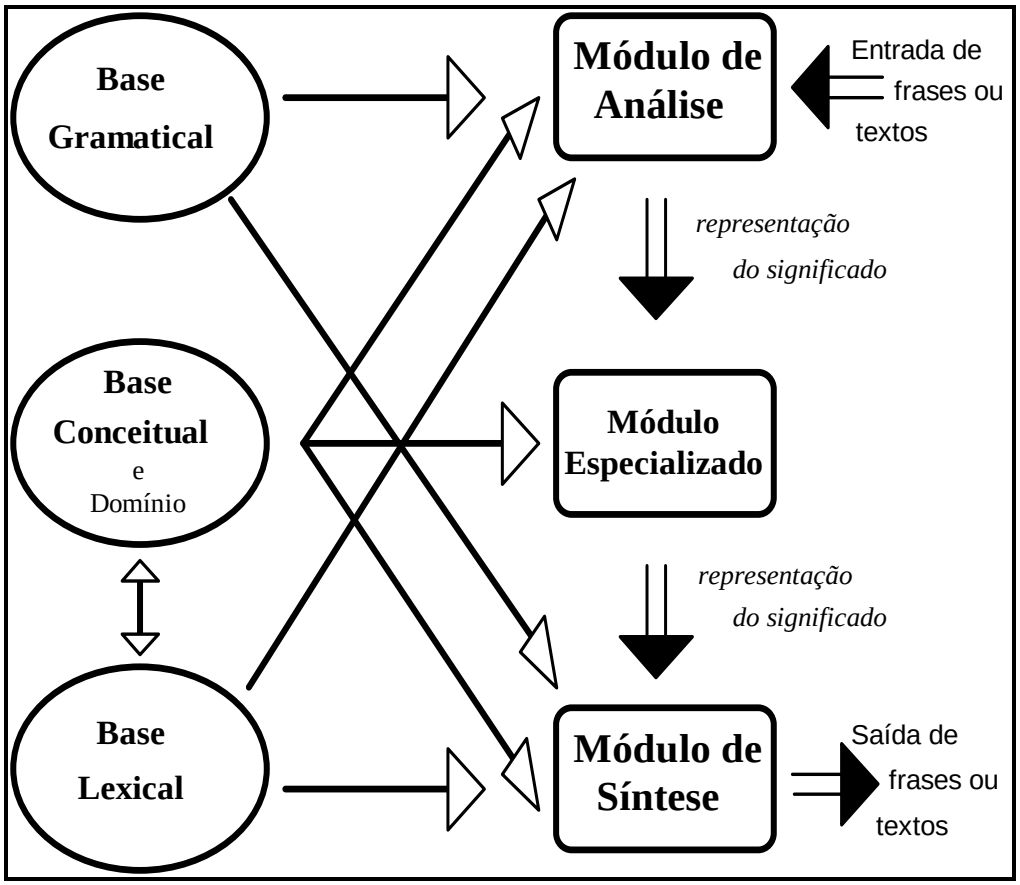
serve como uma espécie de interlíngua entre as línguas naturais e a linguagem da máquina;

- sistemas que utilizam modelos de conhecimento sobre o mundo: o conhecimento de mundo é, em geral, codificado em termos de *frames*, *scripts* ou *redes semânticas*;
- sistemas que incluem informação sobre os objetivos e crenças de agentes: as intenções e crenças são codificadas em termos de *planos*.

Componentes essenciais

Não há exatamente um acordo sobre o número exato de componentes que um sistema de PLN deva possuir. Observa-se, contudo, que dois grupos de componentes intimamente inter-relacionados são tacitamente propostos como imprescindíveis para a implementação de quaisquer sistemas dessa natureza: um grupo de componentes que armazenam as "bases de conhecimento estático" – as bases gramatical, lexical e conceitual –, e um grupo de componentes que processam todas as informações que entram, circulam e saem do sistema – os módulos de processamento, que operam sobre essas bases.

O esquema, a seguir, sugere, com base em Nirenburg *et al.* (*op. cit.*), uma possível arquitetura para um SPLN:



As setas simples (→) representam o fluxo de informações que partem das *bases de conhecimento estático* para os *módulos de processamento*; as setas duplas (⇒) representam as "transformações" sucessivas por que passam as representações; a seta de dupla direção (↔) entre o módulo *lexical* e a base *conceitual* representa a indexação que se estabelece entre os itens lexicais e a estrutura de conceitos.

Com exceção do **Módulo Especializado** (ME)**,** que deve ser projetado para realizar tarefas específicas em função do sistema de que será parte, os demais módulos e as bases de conhecimento estático, em

geral, possuem uma estrutura e um funcionamento padrão, embora os conteúdos possam variar em função da especificidade do sistema.[118]

A **BASE GRAMATICAL** contém a representação das regras sintáticas, ampliadas com equações funcionais, que especificam a forma das projeções sintáticas e funcionais de frases (*cf.* a TLF no quarto capítulo). Essas regras podem ser entendidas como condições de admissibilidade de estruturas sintáticas bem-formadas, condições que servirão de referência para o módulo de análise que, por sua vez, se encarregará da construção das representações sintáticas e funcionais e das projeções semânticas e pragmático-discursivas.

A **BASE LEXICAL** armazena uma coleção de unidades lexicais. Para cada unidade, faz-se necessária a especificação de conjuntos de feixes de traços morfológicos, sintáticos, semânticos e pragmático-discursivos. Além desses elementos de natureza lingüística, cada unidade deverá conter também a especificação de elementos que poderiam ser denominados computacionais como, por exemplo, variáveis, procedimentos, seqüências de instruções e outros objetos tipicamente encontrados em programas computacionais. Com a adição desses elementos "estranhos" à descrição usual de objetos lingüísticos, cada "entrada" do "dicionário computacional", criado especificamente para um SPLN, constitui em si mesma uma espécie de programa computacional, codificado em uma linguagem de programação e, dependendo de sua função sintática ou valor semântico, executa uma série de procedimentos durante o processamento.

---

118 O módulo especializado pode fazer parte, por exemplo, de um sistema tutor, de um sistema de tradução automática, de um sistema especializado, de um sistema acadêmico, e assim por diante (*cf.* os vários tipos de SPLN comentados no segundo capítulo).

A **BASE CONCEITUAL** contém um modelo do mundo físico e conceitual, descrevendo tipos básicos de objetos, eventos, forças, propriedades, relações e atributos em termos de representações hierarquicamente estruturadas, isto é, a sua estrutura lógica consiste em uma rede de unidades conceituais interligadas (*cf.* as redes semânticas no capítulo anterior). O "modelo de mundo" desempenha um papel fundamental no conjunto do sistema, fornecendo definições uniformes para as categorias semânticas básicas, que por sua vez constituem as unidades atômicas utilizadas na descrição dos modelos particulares e do léxico. Além de delimitar a "visão de mundo" simulada pelo sistema, essa base constitui uma espécie de "matriz conceitual" em que os vários domínios específicos encontram-se indexados.

Acoplados a essa base de conceitos gerais, prevêem-se conceitos mais específicos, isto é, conceitos referentes a domínios particulares do conhecimento ou conceitos relacionados a atividades, ou tarefas, para a qual o módulo esteja sendo projetado. Observe que esse recorte torna-se absolutamente necessário, uma vez que criar um "repositório universal de conhecimentos", capaz de armazenar todo o tipo de conhecimento já acumulado pelo homem, é uma tarefa impraticável. Por meio da criação de micromundos e do uso de estruturas como *frames*, *scripts* (DYER *et al.*, 1990), planos, do raciocínio por omissão (do inglês *default reasoning*) e do foco de atenção dos participantes do discurso (*cf.* quarto capítulo) consegue-se restringir tanto o universo do discurso quanto a explosão de inferências plusíveis, decorrentes do processo de interpretação do texto. A representação do contexto situacional e do conhecimento de mundo inclui, basicamente, representações de crenças dos agentes

comunicativos e a sua utilização na construção de modelos computacionais que simulem os atos de fala (*cf.* capítulo anterior).

Nos dois processos complementares de recepção e de produção de frases ou textos, as diferenças intrínsecas ao sentido do fluxo de informação servem de parâmetro para a caracterização dos dois módulos básicos de processamento: o **Módulo de Análise** (MA) e o **Módulo de Síntese** (MS), respectivamente. O MA, projetado para decodificar frases ou textos digitados em um terminal, ou já armazenados no sistema, produz, como resultado do processamento, representações abstratas do significado desses objetos lingüísticos. Essas representações são, na seqüência do processamento, passadas a um ME.

O MA recebe uma frase, ou seqüência de frases, digitada via teclado, e constrói uma representação interna do significado (*cf.* a exemplificação do modelo de representação abstrata do significado do texto no capítulo anterior). Ao executar essa tarefa, e dependendo da sofisticação SPLN de que é parte, esse módulo, utilizando-se das bases gramatical, conceitual e lexical, deverá executar todas ou parte das seguintes análises: morfológica, sintática, semântica, incluindo procedimentos de desambigüização, de determinação de referentes e de elementos pragmáticos como, por exemplo, informações sobre atos de fala, atitudes do locutor e situação da comunicação.

Executadas essas tarefas (ou parte delas), a representação do significado resultante é, então, transferida ao ME, que executa a tarefa específica para a qual foi projetado. Por exemplo, se o ME é parte de um sistema de consulta à base de dados, sua tarefa consiste em interpretar as instruções, selecionar a informação solicitada na base de

dados e, em seguida, transferi-la para o módulo de síntese que, se encarregará de gerar o texto apropriado.

A estrutura interna do módulo de análise é composta de um analisador morfológico, um analisador sintático, "derivado" de modelos de descrição gramatical, um interpretador semântico, que converte frases e textos em representações semânticas abstratas, e um interpretador "discursivo-contextual", responsável pela interpretação da estrutura temática, informacional, pragmático-discursiva e contextual.

Assim, no processo de análise dos enunciados digitados pelo usuário, o sistema, partindo da seqüência de palavras, deve construir e deduzir o seu contéudo proposicional e as prováveis intenções do digitador. De maneira geral, esse processo espelha os estágios de análise ditados pela tradição dos estudos lingüísticos: análise morfológica, sintática, semântica e pragmático-discursiva. Consideram-se elementos dados, isto é, os elementos explicitamente disponíveis para uso do sistema, a seqüência linear de palavras. Logo, a primeira tarefa a ser executada é a análise gramatical da seqüência de palavras que compõe o texto.[119] Durante essa etapa, a representação abstrata do significado do texto vai sendo gradualmente construída. O procedimento de "leitura" recorre a dispositivos dinâmicos de construção de representações que, a todo momento do processo, avaliam hipóteses, fazem previsões múltiplas e, gradativamente, vão expandindo as representações.

O MS opera de modo inverso ao MA. Sua tarefa é receber, do ME, uma representação abstrata do texto e transformá-la em uma seqüência de "frases contextualizadas". Os "textos" produzidos pelo MS

119 Neste caso o sistema já recebe a palavra analisada morfologicamente. Caso contrário, seria preciso proceder à análise morfológica.

podem ter extensões bastante diversas: desde um único sintagma ou uma única frase, em resposta a alguma pergunta específica, passando por seqüências de frases declarativas e interrogativas em diálogos, até parágrafos e páginas inteiras, fornecendo comentários, explicações e definições.

Dependendo de sua sofisticação, ele simplesmente seleciona um texto pré-armazenado no sistema, caso em que não há síntese propriamente dita. Em MEs mais elaborados, tarefas complexas podem ser executadas como, por exemplo, demarcação dos limites de frases, seleção de itens lexicais apropriados ao contexto, utilização de recursos dêiticos e anafóricos, omissão de informação recuperável pelo contexto, simulação de atos de fala, de graus de polidez, do uso de operadores discursivos.

Num sistema de "perguntas e respostas" (*cf*. LEHNERT, 1986; WEBBER, 1990), em que o sistema "dialoga" como o usuário, por exemplo, constrói-se uma estrutura computacional que registra as representações do discurso em andamento e interliga o MA ao MS: a "estrutura de agenda". Essa estrutura é uma espécie de "quadro de avisos" a que ambos os módulos têm acesso e a que ambos recorrem para retirar ou colocar informações.

Além da estrutura de agenda, em que a história do discurso vai sendo registrada, é possível também prever um mecanismo de "controle do discurso" que restringe as ações do MS e coordena as interpretações propostas pelo MA. Os SPLNs projetados com esse grau de sofisticação desempenham, comumente, o papel de uma espécie de "tradutor" entre o usuário e o que se costuma denominar "programa subjacente". O programa subjacente pode ser, por exemplo, um

gerenciador de base de dados ou um sistema especializado em alguma área tecnológica. Nesses casos, o controle do diálogo homem-máquina é exercido pela máquina, que armazena informações detalhadas sobre a especialidade para a qual foi projetada e possui mecanismos inferenciais que permitem deduzir e, até mesmo prever, certos passos do raciocínio do usuário, à medida que a sessão de consultas se desenvolve (*cf*. HENSCHEN, 1990; KUIPERS, 1990).

Assim, o processo de produção automática de um texto inicia-se no interior do programa subjacente, quando deste é solicitada alguma informação pelo usuário. Uma vez iniciado, três tipos de subprocessos são executados:

(i) identificação dos objetivos comunicativos que o enunciado deve atingir;

(ii) planejamento de como os objetivos devem ser alcançados, incluindo a avaliação da situação e dos recursos comunicativos disponíveis;

(iii) realização dos planos em forma de texto.

O primeiro subprocesso restringe-se, em geral, a fornecer algum tipo de informação para o usuário e a induzi-lo a executar alguma ação, ou a desenvolver algum raciocínio. O segundo subprocesso envolve a seleção (ou omissão deliberada) de unidades de informação (conceitos, relações, indivíduos) que devem ou não aparecer no texto e a adoção de uma estrutura ou esquema retórico global (progressão temporal, comparação, contraste, enumeração). O terceiro subprocesso refere-se à textualização dos planos. Este último, portanto, depende de um complexo de conhecimentos: gramatical, pragmático e discursivo.

Esses três subprocessos são interpretados como dois subcomponentes do MS: o subcomponente **estratégico**, que executa os subprocessos (i) e (ii), e o subcomponente **lingüístico**, também denominado componente **tático**, que executa o subprocesso (iii).

Antes de começar a construir o texto, o MS toma todas decisões a respeito do objetivo comunicativo que deve simular, da estrutura do contéudo e da forma geral do texto. Feitas as escolhas, o MS procede à linearização dessas estruturas hierárquicas, isto é, passa a construir a expressão superficial do texto. Assim, o processo de contrução do texto segue a seguinte seqüência:

- a identificação dos objetivos comunicativos precede a escolha e o detalhamento dos conteúdos proposicionais;
- o planejamento da estrutura retórica que envolve a mensagem precede a construção das estruturas sintáticas;
- o contexto sintático de uma palavra é fixado antes da escolha da forma morfológica.

Embora esse procedimento descendente de construção seja lingüisticamente motivado, o processo de geração do texto parece ser mais bem definido e dissecado em termos dos procedimentos de planejamento (*planning*). Dessa perspectiva, há dois problemas centrais: organizar o conteúdo a ser expresso e escolher, entre diferentes formas de expressão desse conteúdo, aquelas apropriadas ao contexto. Em outras palavras, o MS precisa, primeiro, planejar o que expressar lingüisticamente para, depois, decidir como expressar. Esses dois problemas têm sido parcialmente solucionados, empregando-se, respectivamente, a metodologia desenvolvida nos sistemas de

planejamento (*planners*) e os algoritmos desenvolvidos para as gramáticas de unificação funcional (a GLF, por exemplo).

Assim, "o problema", que o MS deve resolver, resume-se no seguinte: como realizar um objetivo comunicativo na presença das restrições e das limitações impostas pelos recursos lingüísticos de que dispõe. Decorre, então, que a maior tarefa passa a ser a tomada de decisões:

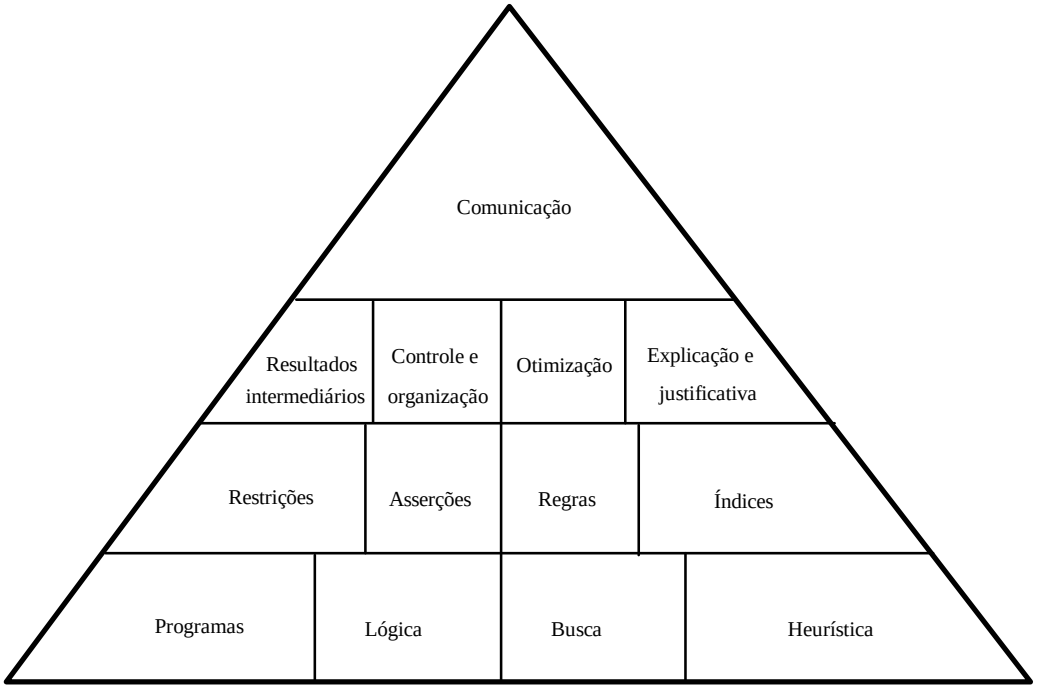
- escolher as palavras;
- selecionar construções sintáticas;
- prever as restrições subseqüentes resultantes das escolhas já feitas.

Tanto a gramática quanto o léxico de uma língua passam a constituir, ao mesmo tempo, **recursos**, pois definem os elementos disponíveis para a construção de textos, e **restrições**, pois as interdependências entre esses elementos obedecem a restrições de co-ocorrência.

Há que se levar em conta que os elementos explicitamente disponíveis para o MS são as representações das "intenções" que deverão simular e os meios de realizá-las. Além disso, à medida que vai construindo as frases, o MS dispõe também dos segmentos parciais de texto já construídos, segmentos que fornecem a necessária ancoragem contextual para a construção das frases subseqüentes. E mais, o sistema também possui um modelo dos usuários, do tipo de comunicação a ser estabelecido e do tipo de texto a ser construído. Todas essas informações permitem ao sistema escolher entre colocações e construções alternativas, tarefa fundamental para o processo de geração do texto.

O SPLN enquanto um sistema de processamento automático de conhecimentos lingüísticos

Assim, como os "sistemas de conhecimentos" (do inglês *knowledge systems*), os SPLNs dependem de programas, de sistemas de restrições, de estratégias de busca ou pesquisa (do inglês *search strategies*), de regras heurísticas, de resultados intermediários, entre outras técnicas. O esquema abaixo, adaptado de Hayes-Roth (*op. cit.*), resume o conjunto de recursos fundamentais para o desenvolvimento de SPLNs.



Na base desse esquema, estão as técnicas comuns a quase todos os tipos de sistemas: os programas que "alimentam" a máquina, a lógica que possibilita representar proposições e mecanismos de

inferência, as técnicas de busca de informação, e as técnicas heurísticas que reduzem o "tempo" de busca.[120]

No nível seguinte, encontram-se as formas de representação do conhecimento mais utilizadas. Entre os exemplos de restrições, incluem-se "Toda frase deve conter um sinal de pontuação que indique o seu término", "Todo pronome reflexivo deve estar ligado em sua categoria de regência", "Todo sujeito concorda em número e pessoa com o verbo principal da oração". As restrições, de um modo geral, referem-se às condições que devem ser satisfeitas para que um determinado elemento seja considerado lingüisticamente bem-formado. As bases de conhecimento assertivo armazenam, por exemplo, itens lexicais, proposições, regras e estruturas sintáticas e fatos sobre o domínio de discurso. As regras representam formas específicas de conhecimento como, por exemplo, as regras ou princípios de estruturação sintática – "se uma preposição for detectada, então, o próximo constituinte deverá ser um sintagma nominal", as regras que especificam como relacionar fatos sobre o mundo "se o jogo for de futebol, então, esperar tumultos". Os índices servem para designar graus de confiança, validade, intensidade, freqüência, previsibilidade, entre outros, que o sistema deve associar aos dados, regras ou resultados. Por exemplo, se regras de topicalização para uma determinada língua forem muito freqüentes, elas devem ter prioridade máxima (em uma escala de 0,0 a 1,0) em relação às regras não topicalizadas, e o juízo de valor expresso pelas expressões

120 Como vimos durante a descrição de um analisador gramatical, a busca de solução para um determinado problema, por exemplo, pode significar explorar uma série de caminhos alternativos. Lembre-se de que, nos analisadores não-determinísticos, técnicas de busca são fundamentais para os diferentes modos de explorar vários caminhos possíveis. Já nos analisadores determinísticos, técnicas heurísticas são fundamentais. As técnicas heurísticas podem também ser empregadas, quando estamos interessados em testar apenas determinados aspectos de um problema. Se pretendemos montar apenas a estrutura sintática genérica, podemos, por exemplo, ignorar as categorias de tempo, aspecto, modo e voz.

muito bom e *excelente* devem ser especificados, por exemplo, por 0,8 e 1,0, respectivamente.

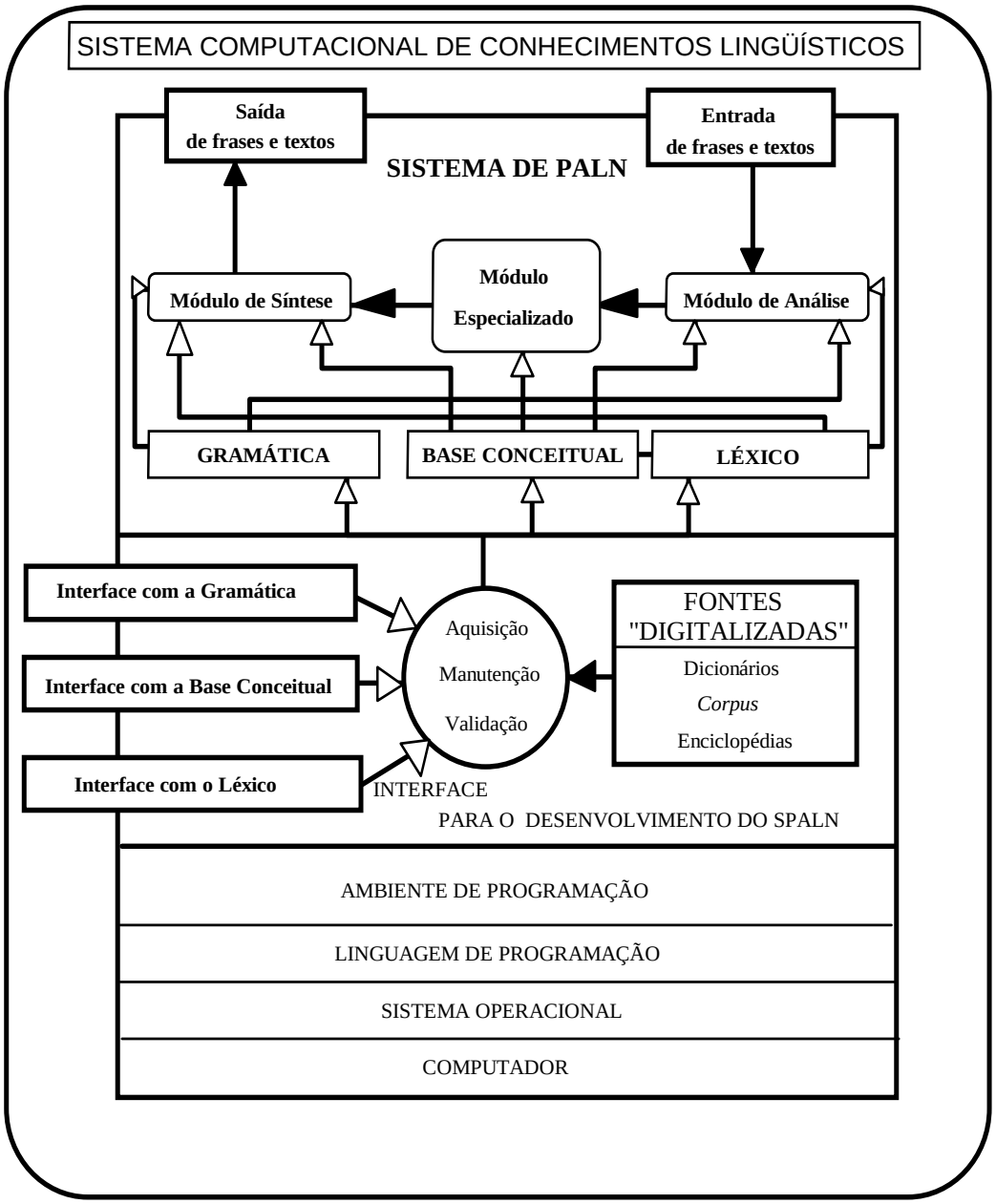
No terceiro nível, os métodos de armazenamento de resultados intermediários são fundamentais no PLN. No processo de construção da "história do discurso" durante um diálogo, é preciso prever uma estrutura que armazene os objetos mencionados, as proposições, os segmentos de discurso já processados, o foco de atenção dos participantes do evento comunicativo, entre outros elementos. Outro momento importante é a construção da representação sintática de frases, que, geralmente, exige o armazenamento de constituintes intermediários que, apesar de já estarem construídos, aguardam para serem integrados a outros constituintes.

Um SPLN organiza e controla suas operações em função de sua própria arquitetura. Essas técnicas são utilizadas para orientar o sistema sobre decisões que precisam ser tomadas. Destacam-se algumas das estratégias que comentei no capítulo anterior: agendas, retrocesso (*backtracking*), quadro-negro, propagação de restrições, programação lógica, estratégia ascendente ou descendente, estratégia por amplitude ou profundidade, entre outras. As técnicas de otimização focalizam o desempenho do sistema. É preciso, nesta fase, examinar se as soluções possíveis são geradas e testadas numa seqüência adequada; se há cálculos redundantes; se as regras são aplicadas com eficiência; se o tempo de processamento está compatível com a tarefa.

As explicações e justificativas referem-se às inferências que o sistema desencadeou para produzir um determinado resultado. O sistema pode ser programado para transformar asserções e regras heurísticas em textos explicativos.

Por fim, um SPLN comunica-se, de um lado, com o usuário, com os especialistas e técnicos e, de outro, com bases de dados, dispositivos de entrada e saída de dados, e outros sistemas. Assim, seu desenvolvimento deve também prever interfaces de comunicação.

Procurei sintetizar no esquema, a seguir, a arquitetura de um "sistema de conhecimentos lingüísticos" ideal que inclui, além dos componentes básicos de um sistema computacional comum, o ambiente para o próprio desenvolvimento de um sistema de processamento automático das línguas naturais:



Dias da Silva, B.C. (1996)

Conclusões e Perspectivas

> *"The computer, like the human mind, has the ability to manipulate symbols in complex processes, including processes that involve decision making based on stored knowledge* […] *By developing and testing computer-based models of language processing that appoximate human performance, researchers hope to understand better how human language works."*
>
> A. Barr & E. A. Feigenbaum (1981: 227)

Apresento uma síntese das principais conclusões a que cheguei com este trabalho, colocando o campo em perspectiva, idealizando trabalhos que poderão contribuir para a proposição de projetos integrados de PLN, incentivando os estudiosos da linguagem a participarem, de fato, de empreendimentos semelhantes e, sobretudo, sugerindo a introdução de estudos dessa natureza no âmbito das Humanidades.

Na essência, esta tese procurou fazer um balanço de uma parcela significativa das pesquisas sobre o PLN e pretendeu mostrar sua particular relevância científica e tecnólogica para o próprio estudo da linguagem humana, e deste para o efetivo desenvolvimento daquelas.

A análise de trabalhos significativos sobre o PLN permitiu a delimitação de alguns contornos de um quadro de referência, um "quase-estado da arte", com o qual pretendi não só ilustrar as tentativas de pesquisadores que acreditam ser possível implementar parcelas do comportamento lingüístico humano no computador, mas também

incentivar a necessária aproximação dos diversos especialistas, cujas contribuições são imprescindíveis para a solidificação do campo. Aos "especialistas da linguagem" cabe enfrentar parte desse grande desafio que, na essência, consiste em criar sistemas de representação, neles projetar complexos de conhecimentos lingüísticos e extralingüísticos e, por fim, codificar as representações resultantes em uma linguagem de programação apropriada.

O processamento da linguagem humana por computadores revela-se, de fato, como um empreendimento fascinante e desafiador, empreendimento que nos coloca como executores de uma tarefa complexa: dissecar e compreender nosso próprio conhecimento lingüístico, a partir de sua interação com outros tipos de conhecimentos, compreender também nossas "ações lingüísticas" e, finalmente, "projetar" esses conhecimentos em modelos formais – tarefa que exige a construção de sistemas de representação diversos e a organização dos mais variados processos.

Das representações, em particular, exige-se precisão, clareza, um alto grau de detalhamento e, sobretudo, compatibilidade com a teoria lingüística que lhe serve de fundamento. De um lado, portanto, as representações devem refletir resultados de investigações lingüísticas e, de outro, precisam ser suficientemente explícitas e estruturadas para que possam ser "transformadas" em programas computacionais que, em última instância, serão os "agentes" que realizarão o processamento automático das línguas naturais.

Em termos concretos, projetar sistemas de PLN com esse grau de sofisticação significa, portanto, criar modelos formais de organização e de representação de informações e procedimentos de

manipulação dessas informações. Desses modelos exige-se um alto grau de "expressividade" para que, por meio deles, seja possível codificar um complexo de informações diversas. Os programas, isto é, a concretização desses modelos formais em estruturas de códigos manipuláveis pela máquina, são, desse modo, concebidos como uma espécie de "forma de expressão" de um complexo de conhecimentos, cuja finalidade específica é instrumentalizar o computador na execução das tarefas de manipulação, interpretação e utilização de parcelas da linguagem humana. Empregando-se o jargão da Inteligência Artificial, cabe aos especialistas em PLN a meticulosa tarefa de projetar as múltiplas "bases de conhecimentos" e, a partir delas, criar os "módulos de processamento" que se encarregarão da manipulação computacional dos "conhecimentos" nelas contidos.

A métrica do grau de sofisticação desses programas é, conseqüentemente, determinada pela quantidade e, principalmente, pela qualidade dos "conhecimentos" que o projetista decide, ou consegue, representar em seu modelo computacional. Os modelos computacionais de PLN mais complexos, criados, por exemplo, para simular um diálogo homem-máquina sobre um determinado assunto, via teclado, exigem, além da representação computacional da estrutura morfossintática e léxico-semântica de frases isoladas, a representação computacional de parâmetros pragmático-discursivos, contextuais e do conhecimento de mundo, elementos essenciais para a simulação de uma interação verbal que se pretenda próxima da interação verbal "natural". Sem a representação desses parâmetros, o modelo não teria elementos para proceder à codificação e à decodificação de expressões referenciais, dêiticas e anafóricas, dos atos de fala e das intenções, atitudes e crenças

dos eventuais usuários. Logo, quanto mais completa e detalhada for a forma de expressão, mais instrumentalizado estará o computador para simular os processos de recepção e produção dos enunciados lingüísticos. Em outras palavras, o grau de "naturalidade" da simulação – até que ponto o "desempenho lingüístico" apresentado pelo sistema pode ser comparado ao desempenho lingüístico humano –, resulta da sofisticação da forma de expressão desenvolvida pelos especialistas.

Vale salientar, porém, que não defendo a criação de máquinas "capazes de se exprimir em linguagem natural" empregando todas as suas potencialidades, como acreditam os defensores mais argutos da Inteligência Artificial. Edward Feigenbaum, um dos inventores dos sistemas especializados, e Marvin Minsky, um dos criadores do campo da inteligência artificial, por exemplo, não só acreditam na viabilidade de se poder criar máquinas "capazes de falar" como vão até muito mais além. Para Feigenbaum, é possível programar um computador a tornar-se um Proust. Ironicamente, chega a dizer que criar programas com essa sofisticação não seria muito "interessante", porque "programas como esses são muito difíceis de calcular, e são portanto muito dispendiosos" (PEPESSIS-PASTERNAK, 1992: 219). Misnky, por sua vez, acredita que os computadores, em alguns anos, não só serão dotados de inteligência, mas principalmente "alcançarão o nível do gênio, e depois o seu poder será incalculável, a tal ponto que teremos sorte se elas resolverem nos conservar como animais domésticos" (*ibid*: 207). Nesta tese, não pretendo, portanto, conceber sistemas de PLN que sejam capazes de produzir e interpretar todas as formas da língua, mas sistemas capazes de cobrir parte das formas, tarefa não menos difícil, quer do ponto de vista computacional, quer do ponto de vista lingüístico.

Diante de empreendimento tão complexo, defendo, acima de tudo, a imprescindível busca de subsídios teóricos em diferentes áreas do conhecimento, estratégia de trabalho que necessariamente exige a aproximação de pesquisadores de áreas de estudo aparentemente tão diversas como Letras, Ciência da Computação e Inteligência Artificial.

Por isso, o estudo do PLN aqui abordado, exige, sobretudo, o desenvolvimento de pesquisas que sejam capazes de ‘‘re-processar’’ contribuições de áreas como filosofia, lógica, psicologia e lingüística, com o objetivo de desenvolver sistemas computacionais em que a comunicação entre o usuário e o computador possa realizar-se com a implementação de linguagens cada vez mais próximas das línguas naturais. Assim, o estudo do PLN não pode prescindir da construção sistemática de um corpo integrado de conhecimentos a partir da aglutinação, inter-relacionamento, complementação e desenvolvimento dos conhecimentos gerados no âmbito das disciplinas matrizes.

Logo, como o corpo de conhecimentos necessários para a implementação de sistemas dessa natureza é oriundo de fontes diversas, sua construção inevitavelmente ultrapassa a fronteira da especialização individual. Por essa razão, para os pesquisadores envolvidos em projetos de PLN, torna-se qualidade imprescindível ser capaz de considerar e partilhar descobertas interdisciplinares. É preciso apostar em trabalhos integrados e mais cooperativos. Não se pode negar uma área do conhecimento para afirmar a outra…

O pesquisador envolvido nesse empreendimento, portanto, além de ter de encontrar soluções para os problemas representacionais colocados pela complexidade da própria linguagem humana – com suas imprecisões, seus sentidos múltiplos, seus jogos de palavras e,

principalmente, ambigüidades e subentendidos contidos em seus enunciados lingüísticos –, precisa também enfrentar dificuldades conjunturais, inerentes às pesquisas interdisciplinares. É por isso que se tornam decisivos, de um lado, a necessária explicitação de um referencial teórico-metodológico mínimo que passe a servir de norte para pesquisas integradas e, de outro, a difícil adoção de posturas científicas mais cooperativas. Buscam-se, acima de tudo, posturas que reconheçam e valorizem as duas modalidades de produção científica que, há muito, receberam os controvertidos rótulos: "ciência pura" e "ciência aplicada". Além disso, com atitudes mais cooperativas, pesquisadores de áreas diversas só teriam a lucrar, beneficiando-se, como já reconhecia Borba (1978: x), da tecnologia "para abreviar caminhos e conseguir novas descobertas''.

O principal problema a ser equacionado parece ainda ser a construção de um arcabouço teórico e metodológico compartilhado, arcabouço que motive, facilite e instigue a aproximação de estudiosos, cujas especialidades e metas de pesquisa possam ser canalizadas para a proposição e para o desenvolvimento de projetos de PLN integrados, e que também estimulem a criação de recursos computacionais específicos para a elaboração de teorias e modelos de descrição lingüística mais explícitos e precisos. Entretanto, a precária troca de experiências, constatada entre os cientistas em geral e, em particular, e, sobretudo, entre os lingüistas e os especialistas das "exatas", pode ser apontada como um dos principais obstáculos que vêm impedindo a realização desse ideal.

Esses desencontros acabam por enfraquecer o embasamento teórico de grande parte dos sistemas de PLN e,

conseqüentemente, por comprometer a sofisticação da simulação, porque, em geral, muitos investigadores envolvidos em projetos de PLN subestimam a complexidade da própria linguagem humana, por desconhecê-la, ou por não terem encontrado modelos lingüísticos que lhes mostrassem um caminho, ou ainda, por não poderem contar com modelos lingüísticos suficientemente explícitos, completos e, de fato, computacionalmente operacionalizáveis.

É fato que existem iniciativas. Além de isoladas, porém, elas norteiam-se por objetivos diversos – de um lado, há lingüistas, tentando empregar recursos da ciência da computação e da informática em suas pesquisas eminentemente lingüísticas e, de outro, há cientistas da computação, matemáticos e lógicos, tentando montar e operacionalizar sistemas de PLN lingüisticamente não motivados, ou absolutamente desprovidos de qualquer valor científico. Além disso, há distorções que precisam ser corrigidas. Bailin & Levin (*op. cit.*: 9), por exemplo, lamentam o fato de que muitos lingüistas, quando se envolvem em projetos computacionais, por incrível que pareça, acabam por ignorar os resultados teóricos já produzidos pela teoria lingüística.

Não são poucos os pesquisadores que vêm advertindo que trabalho cooperativo, envolvendo lingüistas, cientistas da ciência da computação e da inteligência artificial, é condição fundamental para o desenvolvimento de sistemas de PLN. Conclamar investigadores a contribuírem com suas parcelas de conhecimento especializado é condição essencial para que possamos tentar “decifrar o grande enigma das esfinges de nossa época”.

Um trabalho cooperativo entre lingüistas e cientistas da computação, por exemplo, poderia ser realizado na construção de parte

de um dos módulos previstos na arquitetura do sistema de PLN, sugerida neste trabalho – a construção de um analisador gramatical é uma tarefa que pressupõe a construção de um léxico e de uma gramática, com seus princípios, suas regras e suas categorias. Essa tarefa eminentemente lingüística, por sua vez, fornece tanto os elementos constitutivos do analisador gramatical quanto os fundamentos e princípios lingüísticos, que irão complementar e nortear a tarefa computacional de construção de algoritmos e da subseqüente codificação do analisador gramatical em uma linguagem de programação.

Aos lingüistas, especialistas nos estudos da linguagem humana por excelência, caberia um alerta: tornarem-se mais sensíveis às pesquisas que envolvem o tratamento computacional das línguas. Assim como Benveniste trouxe para a lingüística o estudo científico do discurso, por que não fazermos o mesmo com o estudo científico do PLN?

A importância das pesquisas científicas nesse campo ganha destaque ainda maior, quando John Searle, um dos filósofos mais avesso às pesquisas sobre inteligência artificial, acaba por reconhecer o papel que o computador pode desempenhar nas pesquisas lingüísticas. Searle, no Prefácio da tese *Reference and Computation*, de Amichai Kronfeld, publicada em 1990, destaca a ‘‘beleza’’ de se trabalhar com computadores (SEARLE, 1990c: xiii):

> ‘‘A tentativa de implementar computacionalmente uma teoria lingüística constitui uma forma rigorosa de testá-la. A beleza de se trabalhar com computadores é que o computador não deixará você encobrir erros, esquivar-se de questões difíceis ou propor teorias ambígüas.’’

No âmbito acadêmico, a concepção e a arquitetura de sistemas de PLN, aqui propostas, além de poderem contribuir para a abertura de novas frentes de investigação interdisciplinar, poderão também servir de matriz, não só para a implementação de programas computacionais instrumentais, voltados para o desenvolvimento modular de teorias lingüísticas ou do próprio PLN, como também para o desenvolvimento de módulos computacionais de PLN com finalidades práticas específicas.[121]

Os programas instrumentais poderiam, por exemplo, ser utilizados para:

- desenvolver e testar, em separado, de modelos teóricos de representação e de processamento automático de, por exemplo, léxicos, regras sintagmáticas, construções sintagmáticas específicas, tempos e aspectos verbais, verbos causativos, modalidades, atos de fala, elementos anafóricos e dêiticos, e assim por diante;
- investigar as possibilidades de produção automática de resumos de textos, de geração automática de perguntas e respostas sobre textos, ou, até mesmo, de classificação automática de tipos de texto, segundo alguma tipologia pré-definida;
- desenvolver modelos de tradução automática.

Já os módulos de PLN poderiam ser integrados a:

---

121 Ressalto, mais uma vez, que os projetos que considero acadêmicos concentram-se em simulações computacionais que não se restrigem exclusivamente ao desenvolvimento de projetos aplicativos. Os projetos acadêmicos deverão privilegiar a elaboração, teste e refinamento dos próprios modelos de análise e síntese de línguas, contribuindo assim para uma compreensão maior dos próprios fenômenos da linguagem.

- sistemas computacionais práticos, servindo de interface a um sistema tutor, a um sistema especializado ou a uma base de dados, ou ainda, constituindo o "módulo de correção ortográfica", acoplado a um processador de textos;
- sistemas de auxílio ao tradutor, contendo dicionários, gramáticas e enciclopédias informatizados.

O potencial de aplicações teóricas e práticas de sistemas de PLN vem confirmar que pesquisas dessa natureza acabam por estimular a proposição e o desenvolvimento de projetos de vanguarda, apontando para uma possibilidade que, sem o computador, jamais poderia ter sido cogitada – a possibilidade de se **transformar teorias em tecnologias**. Talvez estejamos, de fato, presenciando a gênese de uma **tecnologia lingüística** – a criação de simulacros computacionais da linguagem verbal…

Acredito que o incentivo a e a promoção de pesquisas sistematizadas sobre o PLN, em nossos meios acadêmicos, significam, sem dúvida, abrir novas perspectivas de pesquisas interdisciplinares, pesquisas que podem contribuir não apenas para o avanço do campo da teoria lingüística em geral, mas, principalmente, para o desenvolvimento de tecnologias lingüísticas que podem afetar, de modo decisivo, nossa indústria de informática que, apesar de estar bastante defasada em relação aos países mais desenvolvidos, encontra-se, hoje, em evidente fase de expansão.

Embora tenha enfatizado o lado mais acadêmico dos projetos sobre o PLN, não ignoro seu lado tecnólogico. É importante

lembrar que os aspectos mais tecnológicos dos projetos de PLN encontram ressonância na indústria de informática atual. Devido à natureza "criptográfica" das máquinas, essa indústria necessariamente precisa investir em formas de colocar os computadores mais próximos dos usuários e que, ainda hoje, para utilizá-los, têm que forçosamente ‘‘digerir’’ manuais, decorar combinações de teclas e comandos, prostrarem-se diante das complicadas linguagens de programação ou se adaptarem à "linguagem dos menus". Aproximar os usuários dos computadores significa, acima de tudo, tornar a comunicação homem-máquina mais acessível, principalmente ao usuário não especializado em computação que, na maioria das vezes, afasta-se da máquina por considerá-la ou complexa demais, ou idiota, ou absolutamente dispensável. Facilitar, portanto, a comunicação entre o computador e o usuário, já ‘‘iniciado’’ ou não no universo da informática, implica investigar meios de fazer com que "a máquina do século" comunique-se em linguagem natural.

Em suma, esta pesquisa pretendeu apresentar um conjunto de possibilidades, um estudo de um pesquisador que, como muitos outros, acredita na viabilidade de implementação de programas computacionais capazes de nos ajudar a compreender melhor a própria linguagem humana; estudo que pretendeu apontar um caminho que busca a realização de pesquisas científicas e tecnológicas, [i] investindo na interdisciplinaridade, [ii] mostrando novas perspectivas para a análise dos fatos lingüísticos e [iii] apostando na implementação de sistemas computacionais com potencialidades de não só otimizar a comunicação homem-máquina, mas também instrumentalizar o pesquisador na sua

tarefa de investigação dos graus de operacionalização e consistência de teorias e modelos lingüísticos através de simulações.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMSON, H. & DAHL, V. (1989). ***Logic grammars***. New York, Springer-Verlag.

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

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

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

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

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

ANDREWS, A.D. (1989). ‘‘Lexical Structure.’’ *In*: F. Newmeyer (ed.). ***Linguistics: the cambridge survey I***. Cambridge, Cambridge University Press, pp. 60-88.

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

AUSTIN, J.L. (1962). ***How to do things with words***. New York, Oxford University Press.

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

BAILIN, A. (1989). ‘‘Introduction to intelligent computer-assisted language instruction.’’ ***Computers and The Humanities***, **23**, 3-11.

_____ *et al.* (1988). ‘‘The Use of Natural Language Processing in Computer-Assisted Language Instruction. ***Computers and The Humanities***, **22**, 99-110.

BAILIN, A. & LEVIN, L.S. (1989). ‘‘Introduction: Intelligent Computer-Assisted Language Instruction.’’ ***Computers and The Humanities***, **23**, 1-2.

BAKER, M. (1988). ***Incorporation: a theory of grammatical function changing***. Chicago, The University of Chicago Press.

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

BARR, A. & FEIGENBAUM, E.A. (1981). ***The handbook of artificial intelligence I***. Stanford/Palo Alto, Heuris Tech/Kaufmann, pp. 1-17.

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

BARWISE, J. & PERRY, J. (1983). ***Situations and attitudes***. Cambridge, Mass., The MIT Press, pp. 27-45.

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

BERWICK, R.C. (1985). ***The acquisition of syntactic knowledge***. Cambridge, Mass., The MIT Press.

_____ (1987). ''Intelligent Natural Language Processing: Current Trends and Future Prospects.'' *In*: W.E.L. Grimson & R.S. Patil (eds.). ***AI in the 1980s and beyond: an MIT survey***. Cambridge, Mass., The MIT Press.

BIDERMAN, M.T.C. (1978). ***Teoria lingüística: lingüística quantitativa e computacional***. Rio de Janeiro, Livros Técnicos e Científicos.

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

BOBROW, D.G. (1968). "Natural Language Input for a Computer Problem-Solving System." *In*: M. Minsky. ***Semantic information processing***. Cambridge, Mass., MIT Press, pp. 146-226.

_____ (1986). "GUS: A Frame Driven Dialog System." *In*: B.I. Grosz *et al.* (eds.). ***Readings in natural language processing***. Los Altos, Morgan Kaufmann, pp. 595-604.

BONISSONE, P. (1990). "Plausible Reasoning." *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 854-63.

BORBA, F. S. (1978). ''Apresentação.'' *In*: M.T.C. Biderman. ***Teoria lingüística: lingüística quantitativa e computacional***. Rio de Janeiro, Livros Técnicos e Científicos, p. x.

_____ (1984). ***Introdução aos estudos lingüísticos***. 8. ed. São Paulo, Ed. Nacional.

_____ (coord.) (1991). ***Dicionário gramatical de verbos***. 2. ed. São Paulo, Editora UNESP.

BOSCH, P. (1983). ***Agreement and anaphora***. New York, Academic Press.

BOTT, M.F. (1976). "Lingüística Computacional." In: LYONS, J. (org.) (1987). ***Novos horizontes em lingüística***. São Paulo, Cultrix/Edusp. cap.12, p.206-18.

BRACHMAN, R.J. (1979). ''On The Epistemological Status of Semantic Networks.'' *In*: N.V. Findler (ed.). ***Associative networks***. New York. Academic Press, 3-50.

BRACHMAN, R.J. & LEVESQUE, H.J. (1985). ***Readings in knowledge representation***. San Mateo, Morgan Kaufmann

BRATKO, I. (1986). ***Prolog programming for artificial intelligence***. Workingham, Addison-Wesley.

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

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

_____ (1987). ‘‘LFG User's Manual.’’ ms., 20 p.

_____ (1988). ‘‘The Design of Grammar.’’ ms., 21 p.

BRESNAN, J. & KANERVA, J.M. (1988). ‘‘Locative Inversion in Chichewa: A Case of Factorization in Grammar.’’ ms., 61 p.

BRISCOE, E.J. (1990). ‘‘Speech Understanding.’’ *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 1076-1083.

BRUCE, B. & MOSER, M.G. (1990). “Case Grammar.” *In*: E. Shapiro (ed.) . ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 333-9.

CAMARÃO, P.S.B. (1989). ***Glossário de informática***. Rio de Janeiro, Livros Técnicos e Científicos.

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

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

CHARNIAK, E. (1973). ‘‘Jack and Jane in Search of a Theory of Knowledge.’’ *In*: ***Advanced Papers from The Third International Joint Conference on Atificial Intelligence***, Stanford, pp. 337-343.

_____ (1979). ‘‘Ms. Malaprop, a Language Comprehension Program.’’ *In*: D. Metzing (ed.). ***Frame conceptions and text understanding***. Walter de Gruyter, Berlin, pp. 62-78.

CHIERCHIA, G. & McCONNELL-GINET, S. (1990). ***Meaning and grammar***. Cambridge, Mass., The MIT Press.

CHOMSKY, N. (1957). ***Syntactic structures***. Haia, Mouton.

_____ (1965). ***Aspects of the theory of syntax***. Cambridge, Mass., The MIT Press.

_____ (1981). ***Lectures on government and binding***. Dordrecht, Holanda, Foris.

_____ (1982). ***Some concepts and consequences of the theory of government and binding***. Cambridge, Mass., The MIT Press.

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

_____ (1988). ***Language and problems of knowledge: The Managua Lectures***. Cambridge, Mass., The MIT Press.

_____ (1989). ‘‘Some Notes on Economy of Derivation and Representation.’’ *In*: R. Freidin (ed.)**.** ***Principles and parameters** In **comparative grammar.*** Cambridge, Mass., The MIT Press.

_____ (1992). ‘‘A Minimalist Program for Linguistic Theory.’’ ms., 25 p.

CHOMSKY, N. & LASNIK, H (1991). ‘‘Principles and Parameters Theory.’’ [A ser publicado em J. JACOBS, A.von *et al.* (eds.). ***Syntax: an international handbook of contemporary research***. Berlin, Walter de Gruyter.]

CINQUE, G. (1990). ***Types of A-bar-dependencies***. Cambridge, Mass., The MIT Press.

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

CLOCKSIN, W.F. & MELLISH, C.S. (1987). ***Programming in prolog***. Berlin, Springer-Verlag.

CONTE, M.E. (1989). ‘‘Coherence in Interpretation. Comments on Dieter Vieweger's Paper.’’ *In*: W. Heydrich *et al.* (eds.). ***Connexity and coherence***. Berlin, Walter de Gruyter, pp. 275-82.

COOK, V.J. (1988). ***Chomsky's universal grammar: an introduction***. Cambridge, Mass., Basil Blackwell.

COOPER, R. *et al.* (1990). ***Situation theory and its applications***. Chicago, The University of Chicago Press.

COVINGTON, M.A. *et al.* (1988). ***Prolog programming in depth***. Glenview, Scott, Foresman & Co.

CRUSE, D.A (1986). ***Lexical semantics***. Cambridge, Cambridge University Press.

CULLINGFORD, R.E. (1981). ‘‘SAM.’’ *In*: R. Schank & C. Reisbeck (eds.). ***Inside computer understanding***. Hillsdale, NJ, pp. 75-89.

DALRYMPLE, M.E. *et al.* (1992). ‘‘Relating Projections.’’ ms., 50 p.

DAVIS, E. (1990). “Commonsense Reasoning.” *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 833-40.

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

_____ (1992). ‘‘A Note on Null Objects in Brazilian Portuguese.’’ ***Research Project Report***: CMU-LCL-92-1, Pittsburgh, PA, 41 p.

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

_____ *et al.* (1985). ***Natural language parsing***. Cambridge, Mass., The MIT Press.

DUBOIS, J. *et al.* (1978). ***Dicionário de Lingüística***. São Paulo, Cultrix.

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

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

ENKVIST, N.E. (1989). ‘‘From Text to Interpretability: A Contribution to the Discussion of Basic Terms in Text Linguistics.’’ *In*: W. Heydrich *et al.* (eds.). ***Connexity and coherence***. Berlin, Walter de Gruyter, pp. 369-82.

EVANS, D.A. (1990). ***Areas of CL Research at CMU***. Pittsburgh***:*** Carnegie Mellon University, 3p. (Mimeo).

EVANS, D.A. *et. al.* (1991). ‘‘A Summary of the CLARIT Project.’’ ***Research Project Report***: CMU-LCL-91-2, Pittsburgh, PA, 12 p.

EVANS, D.A. & LEVIN, L.S. (1990). ‘‘Toward The Design of Natural-Language Intructional Systems: Notes from ALICE.’’ *In*: ***Working Notes of the 1990 AAAI Symposium on Knowledge-Based Environments for Learning and Teaching, Stanford University***, pp. 58-62.

FARGHALY, A. (1989). ‘‘A Model for Intelligent Computer Assisted Language Instruction (MICALI).’’ ***Computers and The Humanities***, **23**, 234-50.

FÁVERO, L.L. E & KOCH, I.G.V. (1983). “A Lingüística Textual.” *In*: L.L. FÁVERO & I.G.V. KOCH. ***Lingüística textual: introdução***. São Paulo, Cortez, pp. 11-25.

FILLMORE, C. J. (1968). ‘‘The Case for Case.’’ *In*: E. Bach & R. Harms (eds.) (1968). ***Universals in linguistic theory***. New York, Holt, Rinehart & Winston, pp. 1-88.

_____ (1977). ‘‘The Case for Case Reopened.’’ *In*: P. Cole & J. Sadock (eds.). ***Syntax and semantics 8: grammatical relations***. New York, Academic Press, pp. 59-81.

FODOR, J.D. & FRAZIER, L. (1980). ‘‘Is The Human Sentence Parsing Mechanism an ATN.’’ ***Cognition***, **8**, 417-59.

FORD, M., BRESNAN, J.W. & KAPLAN, R.M. (1982). "A competence-based theory of syntactic closure". *In*: J. Bresnan (ed.). ***The mental representation of grammatical relations***. Cambridge, Mass., The MIT Press, pp. 727-96.

FRAZIER, L. (1989). ''Grammar and Language Processing.'' *In*: F. Newmeyer (ed.). ***Linguistics: the Cambridge survey II***. Cambridge, Cambridge University Press, pp.15-34.

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

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

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

GARDNER, A. *et al.* (eds.) (1981). ''Understanding Natural Language.'' *In*: A. Barr & E.A. Feigenbaum (eds.). ***The handbook of artificial intelligence***. Los Altos, William Kaufmannn, pp. 224-321.

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

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

GEVARTER, W.B. (1984). ***Artificial intelligence, expert systems, computer vision, and natural language processing***. Park Ridge, NJ, Noyes.

GREEN, B. *et al.* (1986). "BASEBALL: An Automatic Question Answerer." *In*: B.I. Grosz *et al.* (eds.). ***Readings in natural language processing***. Los Altos, Morgan Kaufmann, pp. 545-49.

GREIMAS, A.J. & COURTÉS, J. (1979). ***Dicionário de semiótica***. São Paulo, Cultrix.

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

GRIMSHAW, J. (1992). ***Argument structure.*** Cambridge. Mass., The MIT Press.

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

GROSZ, B.J. (1977). ''The Representation and Use of Focus in a System for Understanding Dialogs.'' *In*: ***Proceedings of the Fifth International Joint Conference on Artificial Intelligence***, Cambrige, Mass., pp. 67-76.

_____ (1986). ***Readings in natural language processing***. Palo Alto, Morgan Kaufmann.

_____ *et al.* (1983). ‘‘Providing A Unified Account of Definite noun Phrases in Discourse.’’ *In*: ***Proceedings of the Association for Computational Linguistics***, 44-50.

GROSZ, B., JONES, K. & WEBBER, B. (eds.) (1986). ***Readings in natural language processing***. Los Altos, Morgan Kaufmann.

GROSZ, B.J. & SIDNER, C.L. (1986). ‘‘Attentions, Intentions, and The Structure of Discourse.’’ ***Computational Linguistics***, **12**, 175-204.

GRUBER, J. (1965). ***Studies in lexical relations***. Cambridge/Mass., (Tese de Doutorado).

HAEGEMAN, L. (1991). ***Introduction to government and binding theory***. Oxford, Blackwell.

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

HALLIDAY, M.A.K. & HASAN, R. (1976). ***Cohesion in English***. London, Longmans.

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

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

HATIM, B. & MASON, I. (1990). ***Discourse and the translator***. London, Longman.

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

HAYES, P.J. (1979). “The Logic of Frames.” *In*: D. Metzing (ed.). ***Frame conceptions and text understanding***. Walter de Gruyter, Berlin, pp. 46-61.

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

HENSCHEN, L. (1990). “Reasoning.” *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 822-27.

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

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

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

JACKENDOFF, R (1972). ***Semantic interpretation in generative grammar***. Cambridge, Mass., The MIT Press.

_____ (1977). ***X' syntax***. Cambridge, Mass., The MIT Press

_____ (1983). ***Semantics and cognition***. Cambridge, Mass., The MIT Press.

_____ (1990). ***Semantic structures***. Cambridge, Mass., The MIT Press.

JAEGGLI, O. (1982). ***Topics in Romance syntax***. Dorddrecht, Foris.

JAEGGLI, O. & SAFIR, K. (1989). ''The Null Subject Parameter and Parametric Theory.'' *In*: O. Jaeggli & K. Safir (eds.). ***The null subject parameter***. Dorddrecht, Kluwer, pp. 1-44.

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

JOSHI, A.K. (1990). ''Phrase Structure Grammar.'' *In*: E. Shapiro (ed.) . ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 344-51.

KAMEYAMA, M. (1985). ***Zero anaphora: the case of Japanese***. Stanford/Califórnia, Stanford University (Tese de Doutorado), pp. 302.

KAPLAN, R.M. (1989). ''The Formal Architecture of Lexical-Functional Grammar.'' *In*: C.-R. HUANG & K.-J. CHEN (eds.).***Proceedings of ROCLING II***. Taipei, Republic of China, pp. 1-18.

KARTTUNEN, L. & ZWICKY, A.M. (1985). "Introduction." *In*: D.R. Dowty *et al.* (eds.). ***Natural language parsing***. Cambridge, Mass., The MIT Press, pp. 1-25.

KATZ, J.J. (1972). ***Semantic theory***. New York, Harper.

KATZ, J.J. & POSTAL, P. (1964). ***An integrated theory of linguistic descriptions***. Cambridge, Mass., The MIT Press.

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

_____ (1986). ''Algorithm Schemata and Data Structures in Syntactic Processing.'' *In*: B. Grosz, *et al.* (eds.). ***Readings in natural language processing***. Los Altos, Morgan Kaufmann, pp. 35-70.

KAYSER, H. (1989). ''Some Aspects of Language Understanding, Language Production, and Intercomprehension in Verbal Interaction.'' *In*: W. Heydrich *et al.* (eds.). ***Connexity and coherence***. Berlin, Walter de Gruyter, pp. 342-65.

KEMPSON, R. (1977). ***Semantic theory***. Cambridge, Cambridge Univesity Press.

_____ (1988). "Grammar and conversational principles." *In*: F. Newmeyer (ed.). ***Linguistics: the cambridge survey I***. Cambridge, Cambridge University Press, pp.139-63.

KIMBAL, J.P. (1973). ***The formal theory of grammar***. Englewood Cliffs, Prentice-Hall Inc.

KINDERMANN, J. & MEIER, J. (1988). "An Extension of LR-Parsing for Lexical-Functional Grammar." *In*: U. REYLE & C. ROHRER (eds.). ***Natural language parsing and linguistic theories***. Dordrecht, D.Reidel, pp. 131-48.

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

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

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

KUIPERS, B. (1990). "Causal Reasoning." *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 827-32.

KURZWEIL, R. (1990). ***The age of intelligent machines***. Cambridge, Mass., The MIT Press.

LADUSAW, W.A. (1989). ''Semantic Theory.'' *In*: F. Newmeyer (ed.). ***Linguistics: the cambridge survey I***. Cambridge, Cambridge University Press, pp. 89-112.

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

LEECH, G. (1983). ***Principles of pragmatics***. London, Longman.

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

LEHNERT, W.G. (1979). "The Role of Scripts in Understanding." *In*: D. Metzing (ed.). ***Frame conceptions and text understanding***. Berlin, Walter de Gruyter, pp. 79-95.

_____ (1986). "A Conceptual Theory of Question Answering." *In*: B. Grosz *et al.* (eds.). ***Readings in natural language processing*** . Los Altos, Morgan Kaufmann, pp. 651-57.

_____ (1990). "Story Analysis." *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 1090-99.

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

LEVIN, L.S. (1987). ‘‘Toward a Linking Theory of Relation Changing Rules in Lexical-Functonal Grammar.’’ ***Research Project Report***, Center for The Study of Language and Information, CSLI-87-115, Menlo Park, CA, 49 p.

LEVINSON, S.C. (1983). ***Pragmatics***. Cambridge, Cambridge University Press.

LOBATO, L.M.P. (1986). ***Sintaxe gerativa do português: da teoria padrão à teoria da regência e ligação***. Belo Horizonte, Vigília.

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

_____ (1977). ***Semantics 1& 2***. London, Cambridge University Press.

_____ (1979). ***Introdução à lingüística teórica***. São Paulo, Cia. Ed. Nacional-EDUSP

_____ (1981). ***Linguagem e lingüística***. Rio de Janeiro, Zahar.

MANZINI, M.R. (1992). ***Locality***. Cambridge, Mass., The MIT Press.

MARANTZ, A. (1985). ***On the nature of grammatical relations***. Cambridge, Mass., The MIT Press.

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

MARSHALL, G. (1986). ***Linguagens de programação para micros***. Rio de Janeiro, Campus.

MARÍN, F.M. *et al.* (1989). ‘‘El Proyecto EUROTRA en El Marco de La Investigación Sobre Traducción por Computador.’’ ***Lingüística Española Actual***, **11**, 165-78.

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

McCLOSKEY, J. (1989). ‘‘Syntactic Theory.’’ *In*: F. Newmeyer (ed.). ***Linguistics: the Cambridge survey I***. Cambridge, Cambridge University Press, pp.18-59.

McCORD, M. (1990). “Natural Language Processing in Prolog.” In: WALKER, A. *et al.* (1990). ***Knowledge systems in prolog***. Reading, Mass., Addison-Wesley. cap.5., p.337-450

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

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

MINSKY, M. (1975). ‘‘A Framework for Representing Knowledge.’’ *In*: J. Haugeland (ed.). ***Mind design***. Cambridge, Mass., The MIT Press, pp. 95-128.

MOORE, R.C. (1981). ‘‘Problems in Logical Form.’’ ***Proceedings of The Association for Computational Linguistics***, pp. 117-24.

MORENO FERNÁNDEZ, F. (1990). ‘‘Lingüística Informática e Informática Lingüística.’’ ***Lingüística Española Actual***, **12**, 5-16.

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

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

NUTTER, J. (1990). ‘‘Default Reasoning.’’ *In*: E. Shapiro (ed.) . ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 840-8.

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

PEPESSIS-PASTERNAK, G. (1992). “A inteligência artificial: mito ou realidade?.” *In*: _____ . ***Do caos à inteligência artificial***.São Paulo, Ed. UNESP, pp. 191-259.

PEREIRA, F.C. (1985). “A new characterization of attachment preferences.” *In*: D.R. Dowty *et al.* (eds.). ***Natural language parsing***. Cambridge, Mass., The MIT Press, pp. 307-19.

PEREIRA, F. & WARREN, D.H.D. (1980). ‘‘Definite Clause Grammar for Language Analysis - A Survey of The Formalism and A Comparison with Augmented Transition Networks.’’ ***Artificial intelligence***, **13**, 231-78.

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

PERLIS, D. (1990). ‘‘Nonmonotonic Reasoning.’’ *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley, pp. 849-53.

PERLMUTTER, D.M. (1982). “Syntactic Representation, Syntactic Levels, and the Notion of Subject.” *In*: P. Jacobson & G.K. Pullum (eds.). ***The nature of syntactic representation***. Dordrecht, D.Reidel, pp. 283-340.

PERLMUTTER, D.M. & POSTAL, P.M. (1977). “Toward a Universal Characterization of Passivization.” ***Proceedings of the Annual Meeting of the Berkeley Linguistics Society***. Dept. of Linguistics, University of California at Berkely.

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

PERRAULT, C.R. & ALLEN, J.F. (1980). ‘‘A Plan-Based Analysis of Indirect Speech Acts.’’ ***American Journal of Computational Linguistics***, **6**, 167- 82.

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

PRINCE, E.F. (1988). “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.

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

_____ (1989). ‘‘Subjacency and Principle-Based Parser.’’ ms., 58 p.

PRITCHETT, B.L. & REITANO, J.W. (s/d). ‘‘Parsing with On-Line Principles: A Psychologically Plausible, Object-Oriented Approach.’’ ms., 3 p.

PUSTEJOVSKY, J. & BOGURAEV, B. (1991). ‘‘Lexical Knowledge Representation and Natural Language Processing.’’ ***IBM Journal of Research and Development***, **35**, 1-20.

PYLYSHYN, Z.W. *et al.* (1980). “Understanding Natural Language.” In: B.W. ARDEN (ed.) (1980). ***What can be automated***. Cambridge, Massachusetts, The MIT Press. p. 463-74.

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

RAPHAEL, B. (1968). “SIR: A Computer Program for Semantic Information Retrieval.” *In*: M. Minsky. ***Semantic information processing***. Cambridge, Mass., MIT Press. p. 33-145.

RAPOSO, E.P. (1986). “On the Null Object in European Portuguese.” In: JAEGGLI, O. & SAFIR, K. (eds.) (1986). ***The null subject parameter***. Dordrecht, Kluwer.

RAPOSO, E.P. (1992). ***Teoria da gramática. A faculdade da linguagem***. Lisboa, Caminho.

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

REICHMAN, R. (1985). ***Getting computers to talk like you and me***. Cambridge, Mass., The MIT Press.

RÉVÉSZ, G.E. (1983). ***Introduction to formal languages***. New York, Dover.

REYLE, U. & ROHRER, C. (1987). ***Natural language parsing and linguistic theories***. Dordrecht, D.Reidel.

REYTER, R. (1987). “Nonmonotonic Reasoning.” ***Annual Review of Computer Science***, **2**, 147-86.

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

_____ (1985). ‘‘Artificial Intelligence and The Humanities.’’ ***Computers and The Humanities***, **19**, 117-22.

ROBERT, T.L. (1993). ***Informática do quotidiano***. Lisboa, Gradiva.

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

SANDERS, A. & SANDERS, R. (1989). ‘‘Syntactic Parsing: A Survey.’’ ***Computers and The Humanities***, **23**, 13-30.

SCHA, R. *et al.* (1990). ‘‘Discourse Understanding.’’ *In*: E. Shapiro (ed.) . ***Encyclopedia of artificial intelligence***. New York, Wiley. p. 233-45.

SCHANK, R.C. (1982). ‘‘Reminding and Memory Organization: An Introduction to MOPs.’’ *In*: W.G. Lehnert & M.H. Ringle (eds.). **S*trategies for natural language processing***. Hillsdale, NJ, Lawrence Erlbaum. p. 455-93.

SCHANK, R.C. & ABELSON, R. (1977). ***Scripts, plans, goals, and understanding***. Hillsdale, NJ, Lawrence Erlbaum.

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

SCHUBERT, L. & PELLETIER, F. (1982). From English to Logic: Context-Free Computation of ‘Conventional’ Logical Translation.’’ ***American Journal of Computational Linguistics***, **8**, 27-44.

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

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

_____ (1990c). ‘‘Foreword.’’ *In*: A. Kronfeld. ***Reference and computation***. Cambridge, Cambridge University Press. p. xiii-xvii.

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

SHIEBER, S.M. (1986). ***An introduction to unification-based approaches to grammar***. Chicago, The University of Chicago Press.

SIDNER, C. (1979). ‘‘Discourse and Reference Components of PAL.’’ *In*: D. Metzing (ed.). ***Frame conceptions and text understanding***. Walter de Gruyter, Berlin. p. 120-33.

_____ (1983). ‘‘Focusing in the Comprehension of Definite Anaphora.’’ *In*: M. Brady & R. Berwick (eds.). ***Computational models of discourse***. Cambridge, Mass., The MIT Press. p. 267-330.

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

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

SLOCUM, J. (1985). “Machine Translation.” ***Computers and the Humanities***, **19**, 109-116.

_____ (1989). “Machine Translation: practical issues.”*In*: J.A.Campbell & J. Cuena (eds.). ***Perspectives in artificial intelligence***. Chichester, Ellis Horwood. p. 13-38.

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

STERLING, L. & SHAPIRO, E. (1986). ***The art of prolog***. Cambridge, Mass., The MIT Press.

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

TOMITA, M. (1986). ***Efficient parsing for natural language***. Boston, Kluwer Academic Publishers.

TOWNSEND, C. (1990). ***Técnicas avançadas em turbo prolog***. Rio de Janeiro, Campus.

TURNER, R. (1984). ***Logics for artificial intelligence***. Market Cross, Chichester, Ellis Horwood Ltd.

VELDE, R.G. (1989). “Man, Verbal Text, Inferencing, and Coherence.” *In*: W. Heydrich *et al.* (eds.). ***Connexity and coherence***. Berlin, Walter de Gruyter. p. 174-217.

VIEHWEGER, D. (1989). ‘‘Coherence – Interaction of Modules.’’ *In*: W. Heydrich *et al.* (eds.). ***Connexity and coherence***. Berlin, Walter de Gruyter. p. 256-74.

WEBBER, B.(1987). ‘‘So What Can We Talk About Now?’’ *In*: M. Brady & R. Berwick (eds.). ***Computational models of discourse***. Cambridge, Mass., The MIT Press. p. 331-71.

_____ (1990). ‘‘Question Answering.’’ *In*: E. Shapiro (ed.) . ***Encyclopedia of artificial intelligence***. New York, Wiley. p. 814-22.

WEINBERG, A.S. (1989). ‘‘Mathematical Properties of Grammars.’’ *In*: F. Newmeyer (ed.). ***Linguistics: the Cambridge survey I***. Cambridge, Cambridge University Press. p. 416-29.

WESCOAT, M.T. & ZAENEN, A. (s/d). ‘‘Lexical Functional Grammar.’’ ms., 45 p.

WILKS, Y. (1990). ‘‘Machine Translation.’’ *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley. p. 564-71.

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

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

WINSTON, P.H. & HORN, B.K. (1989). ***Lisp***. Reading, Mass., Addison-Wesley.

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

_____ (1978). ‘‘Semantics and Quantification in Natural Language Question Answering.’’ *In*: M. Yovits (ed.). ***Advances in computers***. Vol. 17, New York, Academic Press. p. 2-64.

_____ (1985). “What's in a Link: Foundations for Semantic Networks.” *In*: R. J. BRACHMAN & H.J. LEVESQUE. ***Readings in knowledge representation***. San Mateo, Morgan Kaufmann. p. 218-41.

_____ (1990). ‘‘Augmented Transition Network Grammar.’’ *In*: E. Shapiro (ed.). ***Encyclopedia of artificial intelligence***. New York, Wiley. p. 323-33.