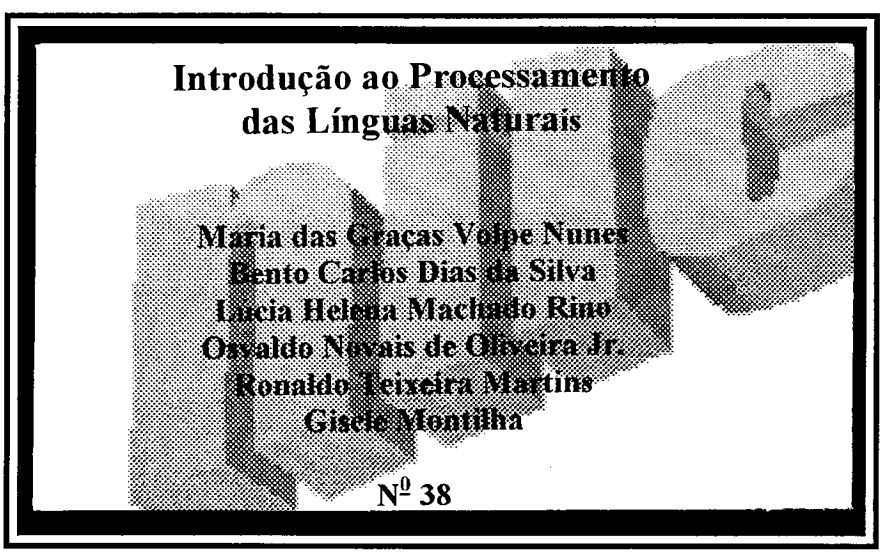
UNIVERSIDADE DE SÃO PAULO



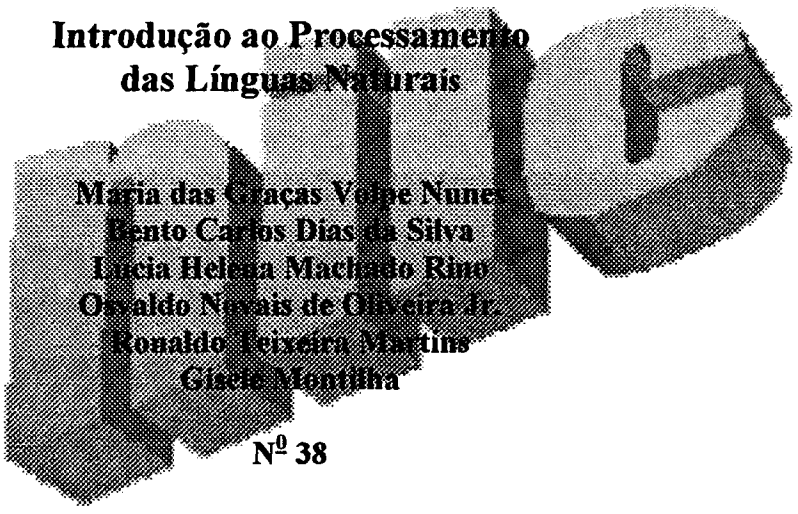
NOTAS DIDÁTICAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2577

Introdução ao Processamento das Línguas Naturais



Maria das Graças Volpe Nunes
Bento Carlos Dias da Silva
Lucia Helena Machado Rino
Osvaldo Novais de Oliveira Jr.
Ronaldo Teixeira Martins
Gisele Montilha

Nº 38

NOTAS DIDÁTICAS DO ICMC

São Carlos
Junho/1999

Esta publicação teve o apoio do Departamento de Ciências da Computação e Estatística do ICMC-USP; da FINEP e da Itautec-Philco S.A. (convênio #88.98.0591-00); da FAPESP (proc.#97/02608-1) e do CNPq (proc.#301365/91-1).

ÍNDICE

INTRODUÇÃO AO PROCESSAMENTO DAS LÍNGUAS NATURAIS

1. Introdução

Este texto pretende introduzir o leitor à área de pesquisa e desenvolvimento em Processamento de Línguas Naturais (PLN). Sem se aprofundar nos mais variados tópicos que serão abordados, o texto tem, antes de tudo, o objetivo maior de motivar o leitor à exploração dessa área e estimulá-lo ao aprofundamento dos tópicos que mais lhe interessam. É intenção dos autores produzir textos que dêem continuidade a esse e, portanto, alarguem o horizonte do leitor que inicia seus estudos nessa área.

O PLN abrange várias e complexas áreas do conhecimento e, por isso, exige que adotemos uma certa perspectiva a fim de traçar uma visão da área. Nesse caso, estaremos focalizando nesse texto apenas o processamento de línguas escritas e, em grande parte das vezes, apenas de textos mono-sentenciais. Além disso, nossas motivações e ilustrações, em geral, referem-se à língua portuguesa escrita no Brasil.

O leitor encontrará, nas seções seguintes, um breve histórico da área de PLN (Seção 2), uma introdução aos diferentes tipos de conhecimento lingüístico para o tratamento de línguas naturais (Seção 3), a apresentação de arquiteturas de sistemas de interpretação e geração de línguas naturais (Seção 4), uma introdução ao processo automático de análise sintática, tão importante na maioria das aplicações de PLN (Seção 5), a apresentação de algumas aplicações desenvolvidas no Núcleo Interinstitucional de Lingüística Computacional, NILC (Seção 6). Finalmente, as referências bibliográficas representam importantes fontes de informações complementares.

2. O Processamento Automático das Línguas Naturais: história e metodologia

2.1 Que idéia é essa?

Desde a sua introdução em nossa cultura, 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 pesquisa que, sem eles, 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, as teorias das lógicas não-clássicas, entre outras.

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 faze-las "entender" instruções, necessárias para a execução de tarefas? A criação de **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.

Outras linguagens de programação, porém, foram sendo criadas; linguagens que, cada vez mais, foram se distanciando dessa representação imposta pela arquitetura do computador e tornando-se mais inteligíveis, pelo menos do ponto de vista humano. Destacam-se, por exemplo, as linguagens *Lisp* e *Prolog*.

Embora a instrução codificada em *Prolog* seja indiscutivelmente muito mais inteligível que as seqüências enigmáticas da linguagem de máquina, ela evidentemente não é uma instrução codificada em língua natural. 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** ou um **syntax error** como resposta.

Cientes dessa inevitável rigidez das linguagens artificiais, 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, facilitando sobremaneira a comunicação entre o usuário e o computador. A questão colocada foi, então: por que não criar "máscaras" que escondam essa maneira primitiva de comunicação? Essa alternativa, hoje, parece ter sido resolvida com grande sucesso. Os computadores, hoje, dispõem de sofisticadas "máscaras"; a "linguagem das interfaces gráficas", com seus menus, ícones e cores, que não só ocultam o que realmente se passa dentro de um computador, como também os transformam em máquinas muito mais atraentes e fáceis de se operar, uma vez que o usuário não precisa mais digitar dezenas de comandos, muitas vezes obscuros e de difícil memorização.

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

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. As preocupações, porém, foram muito mais além. Por que não ousar? Por que não criar meios que instruam o computador a "traduzir" frases e textos de uma língua para a outra?

Questões como essas motivaram os pesquisadores a investigar o **processamento automático das línguas naturais** (PLN). A partir delas, inúmeros "aventureiros" se dispuseram a criar meios para decifrá-lo. Desde então, criar programas computacionais "inteligentes", até mesmo 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, fascinante.

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.

2.2 Um pouco da história

Assim, a amplitude e a heterogeneidade das pesquisas sobre o PLN, somadas à variedade de interesses dos pesquisadores nelas envolvidos e à diversidade de métodos por eles empregada, tornam a sua apreciação histórica uma tarefa difícil, exigindo de seus historiadores o estabelecimento de recortes que acabam por privilegiar determinados fatos em detrimento de outros. Dentre as leituras possíveis, apresentamos aquela em que se resgatam os momentos decisivos que evidenciam a importância da interdisciplinaridade na proposição de soluções para os problemas postos pelo PLN e enfatizamos o papel decisivo da teoria e análise lingüísticas para a consolidação do campo do PLN.

Para isso, tomamos como eixo da exposição a **tradução automática**, que além de ser considerada pela maioria dos autores o marco inicial do uso do computador para a investigação das línguas naturais, permite também apresentar uma síntese da evolução dos estudos nesse campo.

As primeiras investigações institucionalizadas sobre o PLN começaram a ser desenvolvidas no início da década de 50, depois da distribuição de 200 cópias de uma carta, conhecida como *Weaver Memorandum*, escrita por Warren Weaver, então vice-presidente da Fundação Rockfeller e exímio conhecedor dos trabalhos sobre criptografia computacional. Nessa carta, divulgada em 1949, Weaver convidava universidades e empresas, interessados potenciais, para desenvolver projetos sobre um novo campo de pesquisa que ficou conhecido como ''tradução automática'', ''tradução mecanizada'' ou simplesmente MT (abreviação do inglês "Machine Translation").

Tal documento, embora fosse de caráter predominantemente estratégico, já continha as primeiras preocupações teóricas e metodológicas sobre alguns aspectos importantes que deveriam ser contemplados 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 lingüísticos. Essas

diretrizes, entretanto, não estavam no centro das discussões dos projetistas de sistemas de PLN da época. Para eles, traduzir não era diferente de decifrar códigos. A criptografia – técnica que hoje sabemos ser absolutamente inadequada ao tratamento computacional das línguas humanas – era a única ferramenta de que dispunham para criar os programas tradutores.

Nos dois primeiros anos após a divulgação da carta de Weaver, porém, as pesquisas sobre tradução automática passaram a ser levadas a sério em várias instituições importantes como, por exemplo, no Instituto de Tecnologia de Massachusetts (MIT), a Universidade da Califórnia, na Universidade de Harvard e na 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 lingüístico.

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 protótipo 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 itens lexicais de um número maior de línguas. Os resultados alcançados, entretanto, foram muito aquém do esperado pelas agências financiadoras.

O segmento "traduzido" mecanicamente do russo para o inglês, reproduzido a seguir, é suficiente para ilustrar a má qualidade da tradução gerada pelos primeiros sistemas de tradução automática da época.

> (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.

Esses sistemas simplesmente listavam as várias possibilidades de tradução literal de cada palavra encontrada no texto de origem. Nenhuma tentativa de análise sintática era cogitada. Assim, a grande maioria das "traduções automáticas" não só eram de péssima qualidade, como também exigiam constantes revisões por parte de tradutores humanos. Há que se ressaltar que o 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 ao seu prestígio acadêmico e 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, do histórico relatório elaborado pelo Comitê Assessor de Processamento Automático das Línguas Naturais (*Automatic Language Processing Advisory Committee - ALPAC*). Esse relatório, que continha uma avaliação negativa do nível das pesquisas até então produzidas, concluía que, até aquele momento, não só não se havia conseguido executar a tradução automática de texto científico algum, como também não se havia vislumbrado perspectiva alguma para esse 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 americanas e britânicas reduziram drasticamente seus incentivos. O reflexo imediato dessa decisão foi o desaquecimento das pesquisas nesse campo e, sobretudo, dos projetos que visavam à criação de sistemas com finalidades comerciais.

Além desse documento fulminante, a maioria dos trabalhos, de fato, não demonstrava fundamentação lingüística, o que também contribuiu para o seu descrédito e, de maneira geral, para 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.

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 do comitê assessor ALPAC acabou por penalizar muitos projetos sérios que caminhavam para o sucesso – isto é, projetos embasados na teoria lingüística, que nessa década já havia alcançado grau significativo de maturidade. Um deles, por exemplo, o protótipo 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:

> 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.

Uma vez desvencilhados de interesses 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, citam-se os sistemas TAUM-METEO, SYSTRAN, ATLAS II , EUROTRA e KBMT, desenvolvidos nas décadas de 70 e 80.

Assim, por causa de experiências bem-sucedidas e, de certa forma, resistindo aos impactos negativos do relatório governamental, vários outros projetos de PLN, acadêmicos e comerciais, 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 acadêmicas, não foi totalmente abalado. Em 1970, um desses estudiosos militantes, 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 mostrava para a comunidade científica que a interação homem-máquina por meio de línguas naturais poderia ser uma realidade.

O sistema proposto por Winograd simulava, sob forma de representação gráfica no monitor do computador, o braço de um robô que manipulava um conjunto de blocos

sobre a superfície de uma mesa, por meio da interpretação de instruções em inglês digitadas no teclado do computador. No monitor, via-se o braço do robô executando o que lhe era solicitado. Com esse programa, Winograd demonstrava para a comunidade acadêmica que, mesmo de modo primitivo, a máquina poderia ser programada para processar uma interação homem-máquina por meio de uma língua natural.

A partir de experiências como essa, o PLN passou a constituir, de fato, um objeto "digno" de ser pesquisado. Conseqüentemente, uma multiplicidade de pesquisas acadêmicas passou a se somar às pesquisas comerciais que dominavam o campo.

Para finalizar este breve histórico, o Quadro 2.1 apresenta uma síntese da evolução dos estudos do PLN em termos do grau de sofisticação lingüística alcançado.

Quadro 2.1 Evolução dos sistemas de PLN

Década de 50: A Tradução automática
- sistematização computacional das classes de palavras da gramática tradicional
- identificação computacional de poucos tipos de constituintes oracionais

Década de 60: 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: Consolidação dos estudos do PLN
- implementação de parcelas das primeiras gramáticas e analisadores sintáticos
- busca de formalização de fatores pragmáticos e discursivos

Década de 80: Sofisticação dos sistemas
- desenvolvimento de teorias lingüísticas motivadas pelos estudos do PLN

Década de 90: Sistemas baseados em "representações do conhecimento"
- desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos lingüísticos e extralingüísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos lingüísticos

2.3. Um leque de aplicações

O levantamento dos trabalhos de PLN que, 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. A seguir, apresentamos os principais tipos de aplicação.

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-instrução, 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. Esses SPLNs são genericamente denominados "sistemas de perguntas e respostas". Exemplos significativos são: BASEBALL – 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 – 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 – que auxilia implementadores de sistemas na criação do próprio SPLN; PLANES e JETS – que, além de se comunicarem com o usuário por meio de frases, possuem um dispositivo adicional que monitora a comunicação entre o usuário e o sistema, permitindo-lhe otimizá-la; LUNAR – que é capaz de interpretar vários tipos de frases durante o processo de consulta a informações sobre a geologia de rochas lunares; e TEXT – que gera textos da extensão de parágrafos como respostas à solicitação de informação sobre os veículos aquáticos da marinha americana.[1]

Sistemas tutores – Há basicamente dois tipos de sistemas de estudo por computador. Os sistemas tradicionais (*computer-aided instruction*) e os sistemas inteligentes (*intelligent computer-aided instruction*). 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 uma espécie de "diálogo

[1] É importante esclarecer que uma simples mensagem de erro, emitida por um programa como resposta a algum tipo de falha do sistema computacional, não pode evidentemente ser 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

socrático" com o aluno, simulando a situação em que aluno e professor discutem tópicos específicos de conteúdo. 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. Alguns exemplos ilustram algumas iniciativas. SCHOLAR é um programa tutor, 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 auxilia o aluno na resolução de problemas de álgebra elementar formulados em inglês. ALICE é um protótipo de sistema tutor de estudos de língua estrangeira. Nele, destacam-se as seguintes características: seu SPLN é capaz de executar análises morfológicas e sintáticas, gerar frases em inglês, francês, espanhol, alemão e japonês e contextualizar os exemplos por meio de textos e imagens.

Sistemas de automação de tarefas administrativas – Esses sistemas auxiliam nas tarefas de rotina de setores administrativos e gerenciais de empresas e instituições. SCHED é um programa capaz de gerenciar agendas de reuniões. GUS 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.

Programação automática – 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 necessários, 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, durante a qual o sistema adquire um modelo dos processos

---

preenchidas por nomes de indivíduos ou objetos diferentes, por exemplo, tais mensagens também não seriam consideradas textos gerados pelo computador.

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.

Sistemas de processamento de textos científicos – Depois de agrupar relatórios de exames radiológicos e convertê-los no formato de uma base de dados, esse tipo de sistema possibilita ao usuário obter informações por meio de perguntas. As informações de entrada e saída do sistema são codificadas em frases que, por sua vez, 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.

Sistemas especializados – 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, embora sejam 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 e conexo. Para preencher essa lacuna, criam-se os sistemas especializados, projetados para utilizar parcelas do conhecimento humano no processo de resolução de problemas. Nesses sistemas, são implementados mecanismos de aquisição, representação e implementação desse conhecimento, o que os torna mais eficientes que os meios mais convencionais de armazenamento, manipulação e transmissão de informações. Projetados com esquemas complexos de decisão, os sistemas especializados 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: DENDRAL – o primeiro sistema especializado, criado para ajudar os químicos a determinar a estrutura molecular; MYCIN – incorpora 400 regras heurísticas escritas em inglês para diagnosticar doenças sangüíneas infecciosas, oferecendo explicações sobre as conclusões ou perguntas por ele geradas; INTERNIST – contém 100.000 julgamentos sobre relações entre doenças e sintomas; HEARSAY-II – combina sistemas especializados múltiplos na tarefa de interpretar segmentos conexos de fala a partir de um léxico contendo 1.000 palavras; e XCOM – incorpora 1.000 regras de implicação lógica para executar a tarefa de configuração dos componentes de um computador VAX.

Tradução automática – 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 relatórios meteorológicos do inglês para o francês, e o projeto 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 do significado para a qual a língua de partida é "traduzida" e, a partir da qual, a língua de chegada é "gerada".

Sistemas acadêmicos – Pesquisadores como Schank e Riebeck, 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 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 de pequenas histórias.

2.4 Questões conceituais e metodológicas

Nesse emaranhado de pesquisas, adotamos a concepção lapidar que Winograd nos deixou. Nela, encontram-se os elementos ideais para o desenvolvimento do empreendimento e, sobretudo, o indispensável embasamento lingüístico:[2]

> "Assumimos que um computador não poderá simular uma língua natural satisfatoriamente se não compreender o assunto que está em discussão. Logo, é preciso fornecer ao programa um modelo detalhado do domínio específico do discurso. Além disso, o sistema possui um modelo simples de sua própria mentalidade. Ele pode se lembrar de seus planos e ações, discuti-los e executá-los. Ele participa de um diálogo, respondendo, com ações e frases, às frases digitadas em inglês pelo usuário; solicita esclarecimentos quando seus programas heurísticos não conseguem compreender uma frase com a ajuda das informações sintáticas, semânticas, contextuais e do conhecimento de mundo físico representadas dentro do sistema."[3]

Além de evidenciar o complexo de conhecimentos e habilidades envolvidos no processo de comunicação verbal, e que precisam estar representados em um sistema de PLN, Winograd nos mostra que pesquisar o PLN pode ser também um modo de investigação acadêmico que pode auxiliar na compreensão dos próprios fatos da língua:

---

[2] (Winograd, 1972)
[3] Grifo nosso.

> "Todo mundo é capaz de compreender uma língua. A maior parte do tempo de nossas vidas é preenchida por atos de fala, leitura ou pensamentos, sem sequer notarmos a grande complexidade da linguagem. Ainda não sabemos como nós sabemos tanto [...] Os modelos [de PLN] são necessariamente incompletos [...] Mas, mesmo assim, constituem um referencial claro por meio do qual podemos refletir sobre o que é que fazemos quando compreendemos uma língua natural ou reagimos aos atos de fala nela codificados."[4]

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

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

A analogia que construímos permite conceber um SPLN 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 esse campo.

4 Grifo nosso.

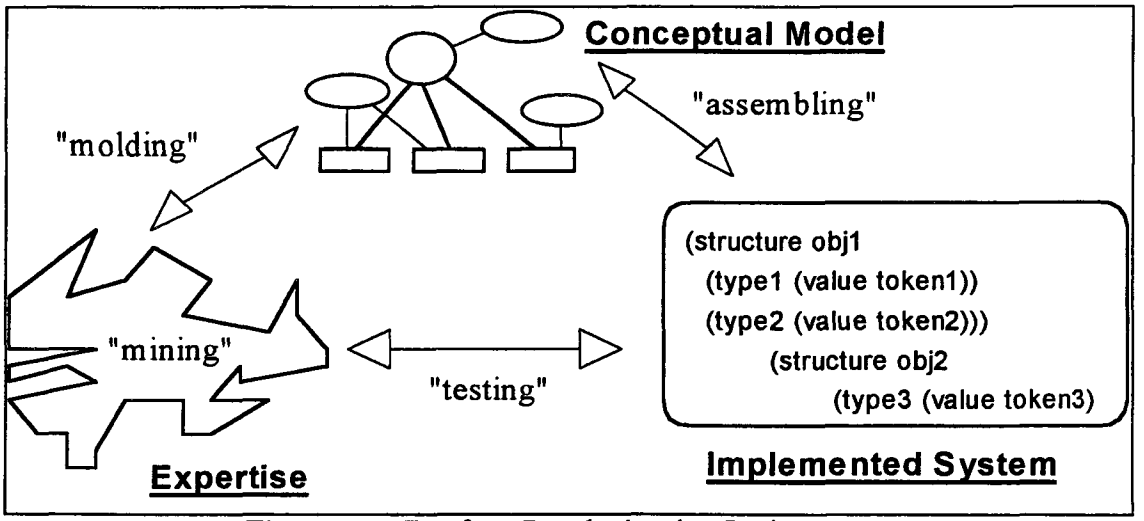


Figura 2.1. Tarefas e Resultados das Explorações

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). A Figura 2.1 ilustra as tarefas previstas e especifica os resultados esperados de cada etapa:[5]

Assim, a explicitação do conhecimento e do 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, 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.

[5] (Dias-da-Silva, 1998)

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

- **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.

Propomos 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. Dessa forma, testes de adequação e de desempenho poderão contribuir para o aprimoramento dos resultados alcançados em cada fase. A dinâmica do processo pode ser visualizada na Figura 2.2.

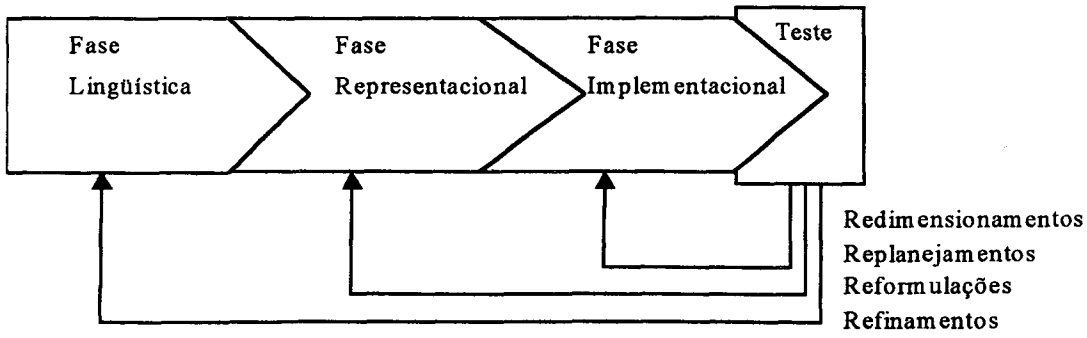


Figura 2.2 A dinâmica do processo de construção de um SPLN

Assim, projetar um SPLN envolve essencialmente (i) especificar, (ii) representar e (iii) codificar sistematicamente um volume 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.

Assim, a grande meta prevista para pesquisas dessa natureza é conseguir projetar e implementar sistemas computacionais avançados em que a comunicação entre o homem e o computador possa se realizar por meio de códigos lingüísticos, e não por meio de instruções e comandos codificados em uma linguagem artificial. 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 "interlocutores e parceiros cibernéticos", capazes de nos auxiliar no planejamento das mais variadas tarefas e, até mesmo, na resolução dos mais difíceis problemas.

Do ponto de vista da pesquisa aplicada, o estudo do PLN deve visar, 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 "pseudolíngua", e não por meio de instruções e comandos convencionais. Nesse sentido, a pesquisa reveste-se de um caráter tecnológico e transforma-se em um objeto cobiçado pela indústria da informática que, cada vez mais, precisa tornar seus produtos menos "enigmáticos" e mais adaptados às necessidades dos seus clientes.

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 um conjunto de programas específicos capazes de executar a complexa tarefa de interpretar e gerar informações contidas em mensagens lingüisticamente construídas. Em outras palavras, estudar o PLN é fornecer subsídios para a implementação de programas computacionais construídos para o fim específico de manipulação de objetos lingüísticos.

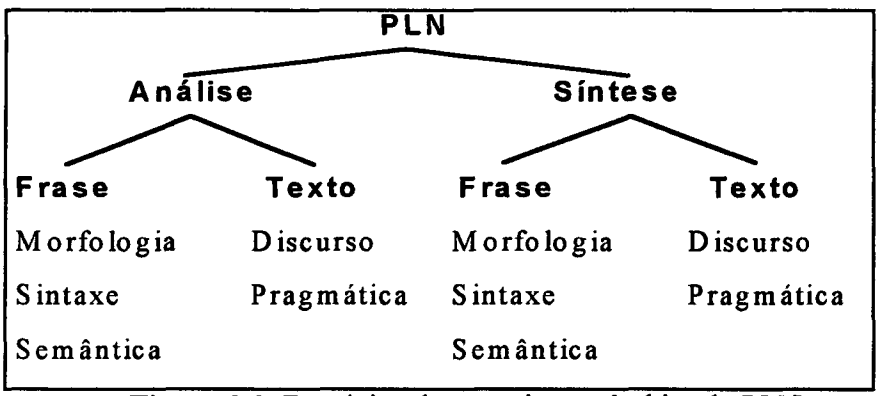


Figura 2.3. Domínios de pesquisa no âmbito do PLN

Além de cumprir objetivos tecnológicos, estudar o PLN significa também desenvolver projetos de caráter acadêmico como, por exemplo, criar modelos computacionais que simulem os processos de produção e recepção de enunciados e textos, ou que sirvam de instrumento no processo de construção e teste dos próprios modelos lingüísticos. Dessa perspectiva, um sistema de PLN passa a ser uma plataforma de trabalho para o desenvolvimento de modelos de análise e descrição lingüísticas, na qual o lingüista auxiliado por projetistas de sistemas de PLN, pode se dedicar à formalização, operacionalização, teste, refinamento e reformulações de seus próprios modelos. A Figura 2.3 apresenta uma síntese dos principais domínios de estudo do PLN.

3. Conhecimento Lingüístico para o Tratamento das Línguas Naturais

No uso cotidiano da língua, a comunicação entre os falantes se dá através de **textos**, embora esses mesmos falantes possuam uma consciência intuitiva das unidades mínimas da língua. Em termos de análise lingüística, pode-se dizer que o texto é a unidade maior na estrutura de uma língua natural, pois reúne em si informações de diversas naturezas que, por sua vez constitui no objeto de estudo de alguns campos específicos na área da Lingüística. A tarefa do lingüista, grosso modo, é identificar e compreender esses segmentos lingüísticos e, a partir daí, apresentar uma descrição do comportamento desses elementos na realização da linguagem verbal. Nesse processo de descrição, então, costuma-se privilegiar os segmentos menores isolados do texto, tornando esse mesmo texto objeto de estudo particular de uma área da Lingüística (Lingüística Textual e Análise do Discurso).

Em PLN o material de entrada do processamento é um texto que deve ser analisado, ou seja, recortado em unidades menores para a compreensão completa dos mecanismos de operação envolvidos em cada dessas unidades. Assim, o PLN recorre àqueles campos específicos da Lingüística, procurando depreender da sua descrição as informações que irão fazer da máquina um instrumento sensível aos fenômenos da língua natural.

Nesta seção vamos focalizar cada um dos tipos de informações lingüísticas que são manipuladas pelo computador no processamento automático da língua.

3.1. A estrutura lingüística

O processador lingüístico costuma recortar o texto em segmentos denominados **sentenças** (S). A análise lingüística automática que opera nesse nível é conhecida como **análise sentencial**, isto é, o tratamento de um texto é promovido de sentença a sentença, sendo ela, portanto, a primeira unidade menor do processamento. Uma sentença pode ser definida como a unidade mínima da *comunicação*, uma vez que se apresenta como um enunciado dotado de expressão completa de sentido. Ela também é conhecida através das denominações de *frase* e *oração*, comumente diferenciadas na Gramática Tradicional da língua portuguesa. No âmbito do PLN, porém, fala-se em sentença para se dirigir aos segmentos organizados das seguintes formas:

1. sentenças constituídas de uma palavra:

Exs.: a. *Atenção!*

b. *Perigo!*

2. sentenças constituídas de um conjunto de palavras no qual se verifica a presença de um verbo (ou locução verbal), ainda que esse verbo esteja oculto:

Exs.: a. *A moça **toca** piano muito bem.* [presença de verbo]

b. *O jogo **tinha terminado**.* [presença de locução verbal]

c. *Ao vencedor, as batatas!* [verbo elíptico]

[Machado de Assis, _____]

3. sentenças constituídas de algumas palavras dentre as quais não há verbo:

Ex.: *Que falsa modéstia, meu Deus!*

Há várias maneiras de descrever as formas pelas quais as sentenças são constituídas na língua, dentre as quais a denominada **análise componencial**. Por esse método de análise uma unidade maior, como a sentença, se constitui de unidades menores imediatamente definidas – os **constituintes** – que se organizam hierarquicamente. Essa disposição hierárquica dos constituintes é a chamada **estrutura interna** da língua e pode ser representada em termos de árvores – *estrutura arbórea* – na qual no topo está a unidade maior (no caso, a sentença), nos níveis intermediários estão elementos sintáticos formativos da sentença (constituintes imediatos) e na base da estrutura, os itens lexicais correspondentes (as palavras). Uma estrutura arbórea simples pode ser ilustrada pelo esquema da Figura 3.1.

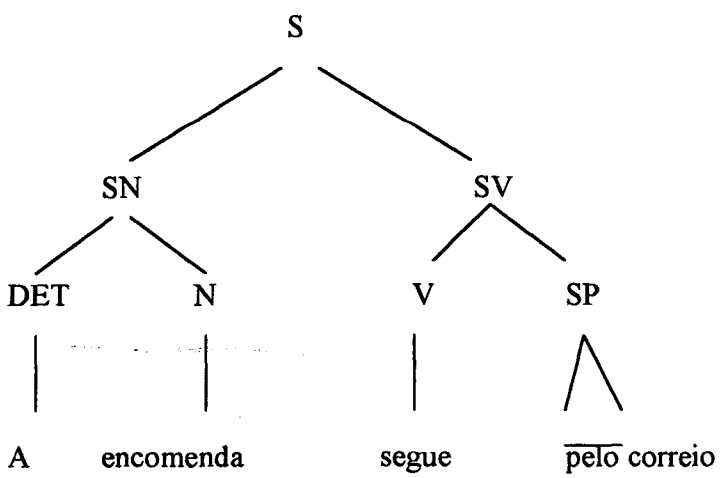


Figura 3.1. Estrutura Arbórea da sentença *A encomenda segue pelo correio*

Nesse exemplo a derivação da estrutura da sentença se constitui dos elementos indicados por SN, SV, SP, DET, N e V que, por sua vez, realizam os itens lexicais da base. Os elementos dos níveis intermediários são a expressão de um conjunto de informações necessário para a composição da sentença. Na estrutura sintática da língua esses segmentos são de dois tipos: sintagmas e categorias gramaticais.

Os **sintagmas** são grupos de palavras organizados em torno de um núcleo sintático que o denomina. Assim, quando o núcleo de um sintagma é um nome (substantivo, adjetivo ou pronome substantivo) falamos de *sintagma nominal* (SN); quando é um verbo (ou locução verbal), há *sintagma verbal* (SV); preposição, *sintagma preposicional* (SP) e advérbio, *sintagma adverbial* (SAdv). As **categorias gramaticais** ou sintáticas, por sua vez, refletem as classes nas quais as palavras da língua são organizadas: determinante (DET), nome (N), verbo (V), advérbio (Adv), preposição (P) e assim por diante.

Para a caracterização de cada um dos sintagmas e das categorias gramaticais da estrutura é necessária a compreensão da função de cada elemento na sentença. Para isso o sistema de processamento é alimentado pelos itens lexicais que carregam toda sorte de informações pertinentes para a sua operacionalização, isto é, aos itens lexicais (palavras da língua) são associadas informações de natureza fonético-fonológica, morfológica, sintática, semântica e pragmático-discursiva. A maneira como essas informações são combinadas para a disposição dos itens lexicais na sentença é dada através das diversas regras de tratamento lingüístico das quais falaremos em outras seções.

É preciso ter claro que esse esquema estrutural que exploramos aqui pode ser aplicado a qualquer nível de descrição da língua. Nesse caso, os nomes dos constituintes categoriais devem ser redefinidos de acordo com os elementos que estão envolvidos no sistema com o qual se pretende operar.

3.2. Os níveis de processamento

Como vimos, as palavras podem ser caracterizadas de diversas maneiras de acordo com o estatuto da descrição lingüística. Isso é devido ao fato de que as palavras possuem propriedades de natureza distinta, refletindo o comportamento que elas adquirem

quando combinadas entre si na atividade comunicativa. Dessa forma, podemos definir uma palavra pelo seu estatuto:

1. *fonético-fonológico*: quando se trata de depreender a identidade sonora dos elementos que constituem a palavra.

2. *morfológico*: quando as unidades mínimas dotadas de significado são isoladas para a compreensão do processo de formação e flexão das palavras.

3. *sintático*: quando a distribuição das palavras resulta em determinadas funções que elas desempenham na sentença.

4. *semântico*: quando o conteúdo significativo da palavra implica relações de natureza ontológica e referencial para a identificação dos objetos no mundo.

5. *pragmático-discursivo*: quando a força expressiva das palavras remete à identificação dos objetos do mundo em termos do seu contexto de enunciação e condições de produção discursiva.

Na maioria dos sistemas de PLN cada um desses níveis de descrição da palavra constitui-se em um **módulo lingüístico**, ou seja, uma etapa do processamento da língua natural. Em cada um desses módulos as informações pertinentes são manipuladas em busca do melhor tratamento lingüístico, seja no reconhecimento seja na produção da língua. É necessário, portanto, que essas informações de que tratam cada um dos níveis de descrição sejam armazenadas junto às formas lingüísticas correspondentes. Nesse caso, a cada palavra do léxico são associadas informações fonéticas – operando com as propriedades sonoras; gramaticais – quando se trata de determinar as suas propriedades morfossintáticas; semânticas – quando as propriedades são da ordem do significado – e pragmático-discursiva – se o conteúdo de expressão revela implicações com o mundo extra-lingüístico (o contexto, o interlocutor, etc.).

Em seguida, apresentaremos de forma sucinta os tipos de informações lingüísticas pertinentes a cada um desses segmentos.

3.3. As informações lingüísticas

(a) <u>Fonético-fonológicas</u>

A análise lingüística denomina de **Fonologia** o estudo do efeito acústico das formas sonoras da língua. Já a **Fonética** ocupa-se da descrição dos sons da fala e das condições pelas quais esses sons são reconhecidos e produzidos pelos falantes de uma língua.

As unidades mínimas do sistema sonoro de uma língua natural podem ser representadas através de formas distintivas do som consideradas **fonemas**. Na língua portuguesa os fonemas realizam os três tipos de sons lingüísticos distintos: *vogais*, *consoantes* e *semivogais*.

Em PLN a representação e operacionalização dos fonemas da língua são particularmente importantes para o processamento envolvido com *síntese de fala*, isto é, a produção pela máquina dos sons produzidos pelo ser humano. Nesse caso, o registro oral é o material de saída (*output*) do sistema de processamento. Quando a máquina opera com o registro escrito, o conhecimento fonético-fonológico ganha importância na determinação dos paradigmas sonoros das palavras da língua, bem como as alterações de timbre e intensidade das palavras motivadas por interferência entre os sons concorrentes do vocábulo.

Alguns fatos lingüísticos podem ilustrar a complexidade do tratamento sonoro das palavras pela máquina:

1. a variação de timbre segundo a caracterização regional das palavras. No Brasil as vogais /e/ e /o/ em posição pré-tônica são pronunciadas de forma "aberta" na região nordeste, ao passo que na região sul e sudeste as mesmas vogais são "fechadas". Exs.: **e**riado; **o**ração.

2. a realização sonora de determinadas formas segundo suas posições na palavra. Nos exemplos a seguir podemos depreender três sons distintos representados pela mesma forma ortográfica: **x**adrez ≠ ê**x**odo ≠ ino**x**.

3. as palavras homófonas (aquelas com mesma forma sonora com significado diferente). Exs.: para / pára; pelo / pêlo.

Esses casos acima demonstram a necessidade do conhecimento das especificidades da cadeia sonora de uma língua a fim de que a ferramenta computacional opere adequadamente com os fonemas. O esforço primordial nesse nível de processamento, porém, está na melhor representação fonética das palavras da língua, assim como a estipulação das restrições fonológicas que cada tipo de som acarreta para o sistema sonoro.

(b) Morfológicas

As palavras da língua também podem ser segmentadas em termos do seu conteúdo significativo. As unidades mínimas dotadas de significado (gramatical ou lexical) são denominadas **morfemas** e se constituem no objeto de estudo da **Morfologia**.

Os morfemas da língua portuguesa podem ser de dois tipos: **gramaticais** – quando se trata de definir os marcadores da flexão das palavras –, e **lexicais** – quando alguns elementos são associados a uma base na formação de novas palavras da língua. Dessa forma, os morfemas gramaticais estão envolvidos no chamado mecanismo de flexão das palavras, que nos nomes identificam os traços de *gênero* e *número* e nos verbos os traços da *conjugação verbal* (pessoa, número, tempo, modo, aspecto e voz). Já os morfemas lexicais ocupam-se do processo de derivação das palavras, dado que a uma base (radical) são associados afixos formando uma palavra nova da língua.

Exs.:

1. *pedra* ⇒ *-a*: morfema gramatical que indica os nomes terminados em "-a".

2. *macaca* ⇒ *-a*: morfema gramatical que indica gênero feminino.

3. *procuramos* ⇒ *-a*: morfema gramatical que indica a primeira conjugação verbal; -mos: morfema gramatical que indica primeira pessoa do plural do presente do indicativo.

4. *operação* ⇒ *-ção*: morfema lexical que indica evento.

5. *incerto* ⇒ *in-*: morfema lexical que indica negação.

Nesse momento do processamento lingüístico são particularmente importantes a determinação exata dos segmentos morfológicos e as relações que eles implicam na definição da palavra. O fenômeno da concordância, por exemplo, é uma dessas relações que exige a presença de certo morfema e não outro no interior da palavra, já que elas não são isoladas no texto. Nesse sentido, a tarefa do investigador da língua é especificar os traços morfológicos pertinentes a cada item lexical, além do mecanismo de operação que esses traços exigem na concatenação das palavras na sentença.

(c) Sintáticas

A organização das palavras na sentença acarreta a definição desses itens lexicais em termos de suas **funções gramaticais**. Trata-se de reconhecer as regras pelas quais a distribuição das formas são determinadas e esse exercício é o objeto de estudo da **Sintaxe**.

Quando as palavras são combinadas entre si para formar um enunciado dotado de um sentido completo, sua distribuição na sentença não ocorre de maneira aleatória, mas, ao contrário, essa disposição segue **regras** estruturais bastante definidas. Essas regras determinam, por exemplo, o emprego dos pronomes, a aplicação da crase, a realização da concordância. Na manipulação dessas regras, faz-se uso de um conjunto de **categorias** definido em termos da sua função sintática, das quais são exemplos as categorias sujeito, objeto direto, complemento nominal, adjunto adverbial e assim por diante.

Em PLN costuma-se reunir na terminologia das categorias gramaticais as entidades sintáticas de que falamos acima e também as classes gramaticais (substantivo, verbo, adjetivo, pronome, numeral). É uma maneira de identificar as palavras segundo o conjunto gramatical ao qual elas pertencem e, ao mesmo tempo, reconhecê-las na sua distribuição sentencial. A atribuição desses traços sintáticos aos itens lexicais constitui uma primeira etapa do tratamento lingüístico no âmbito da sintaxe. Em seguida, são articuladas as regras sintáticas do tipo que levantamos anteriormente a fim de determinar as sentenças bem formadas de uma língua. Podemos ilustrar esses dois momentos da seguinte forma:

Dada a sentença *Ela foram a padaria hoje cedo*, podemos determinar:

1. os traços sintáticos dos itens lexicais:
    a. *ela* $\Rightarrow$ sujeito, pronome pessoal [singular]
    b. *foram* $\Rightarrow$ verbo [de movimento]
    c. *a padaria* $\Rightarrow$ objeto indireto
    d. *hoje cedo* $\Rightarrow$ adjunto adverbial de tempo
2. as regras para especificação de uma sentença bem formada:
    a. concordância verbal obrigatória entre sujeito e verbo

b. emprego da crase obrigatória em complementos preposicionados do verbo (objeto indireto)

A partir dessas informações é possível reformular aquela construção e definir a seguinte sentença adequada da língua portuguesa: *Elas foram à padaria hoje cedo.*

Outras duas noções sintáticas são particularmente importantes nesse nível de tratamento lingüístico. A primeira delas diz respeito à **estrutura argumental** de algumas palavras da língua, especialmente os verbos. Na representação lexical das palavras, bem como na estipulação das regras de transformação é fundamental que se informe ao sistema o número e o tipo dos argumentos exigidos pelo item lexical. Exs.:

1. *pai* é uma palavra de um argumento, uma vez que essa palavra pressupõe a idéia de que se há o objeto pai, esse objeto é pai de alguém, como na sentença: *Meu pai está atrasado para o almoço.*

2. *gostar* é uma palavra de dois argumentos, sendo que o último deve ser preenchido por um elemento acompanhado de preposição, como na sentença: *Benedito gosta de quiabo com pimentão.*

A segunda noção ocupa-se do **papel temático** dos verbos. Considera-se que os verbos atribuem determinadas características aos seus argumentos que devem ser respeitados na construção de uma sentença. Por exemplo: na sentença a seguir pode-se depreender dois papéis temáticos distintos atribuídos pelo mesmo verbo:

1. Aquele rapaz encontrou uma chave na rua.
   [papel temático de "aquele rapaz": agente]
   [papel temático de "uma chave": paciente]

Em síntese, o processamento sintático não faz uso apenas das informações sintáticas que postula (observe, por exemplo, o traço "singular", "humano" e "animado" presentes nos exemplos acima). Apesar disso a sua autonomia é bastante clara em relação ao módulo morfológico, de um lado e semântico, de outro. Esse fato determina o papel central que a análise sintática desempenha no processamento automático de uma língua e que estaremos explorando melhor em outros momentos.

(d) Semânticas

As relações envolvidas no plano do significado das palavras em busca de alcançarem certo sentido no escopo da sentença é a matéria de investigação da **Semântica**. O significado é inerente ao signo lingüístico e está presente não só na palavra como uma unidade completa, mas nas suas unidades constitutivas. Da mesma forma, fala-se em significado de expressões, de sentenças, enfim, de unidades mais complexas da língua. Grande parte do esforço do tratamento semântico em PLN deve envolver, então, a depreensão das propriedades semânticas dos itens lexicais para a construção de sentenças semanticamente bem formadas da língua.

Nessa tarefa está essencialmente presente a idéia dos **traços semânticos** que apontam para um sentido específico do item lexical, e do **conhecimento ontológico** dos objetos no mundo que devem permitir atribuir às palavras informações complementares de sentido. De fato, estamos falando, de um lado, em termos de traços como "concreto", "humano", "animado" e, de outro lado, de categorias ontológicas como "evento", "ação", "coisa", etc. Nessa perspectiva, na representação lexical procura-se definir as informações semânticas primitivas que encaminhem a interpretação da palavra para determinado sentido. Podemos ilustrar esses primitivos semânticos associados às seguintes palavras:

1.a. *homem* ⇒ entidade, concreta, animada, humana, macho

b. *mulher* ⇒ entidade, concreta, animada, humana, fêmea

2. a. *grávida* ⇒ propriedade, abstrata, humana, [ligada à] fêmea

b. *prenha* ⇒ propriedade, abstrata, não-humana, [ligada à] fêmea

Os maiores problemas encontrados no tratamento automático das palavras no que diz respeito a sua especificidade semântica referem-se às **ambigüidades** do tipo polissemia (ex.: a palavra "cabo") e homonímia (ex.: a palavra "ponto"). O trabalho de investigação dos primitivos semânticos que possam representar adequadamente essa ambigüidade é uma tarefa básica no estudo da semântica lexical, acompanhada das regras léxico-semânticas para a interpretação desse fenômeno lingüístico.

(e) Pragmático-discursivas

Nesse nível de análise lingüística estão em foco as questões, consideradas por muitos estudiosos, do mundo **extralingüístico**. Essa noção é amparada pelo fato de que para além das formas e das estruturas, a língua recupera da situação comunicativa diversos fatores que implicam a determinação de certa compreensão das palavras e sentenças. Todo texto é produzido por certos **interlocutores**, em um tempo e um lugar determinado, o que significa dizer que nenhum texto existe independente dos indivíduos envolvidos na atividade comunicativa e nenhum texto existe sem uma situação de **contexto**. Quando se examina uma construção lingüística procurando essas relações presentes no ato da fala, na verdade procura-se estudar aquilo que é objeto da **Pragmática**.

Em PLN é comum associarem ao ambiente da Pragmática aquilo que constitui objeto de estudo da Análise do Discurso por também estar compreendido no mundo extralingüístico. Dessa forma, esse componente do processamento acumula o tratamento de informações mais densas que estão na ordem do discurso: as condições de produção e formação discursiva.

Através da noção de **formação discursiva** quer-se indicar o fato de que os enunciados de uma língua materializam certa *ideologia*, isto é, a ideologia presente no discurso permite ao falante proferir um enunciado e não outro na língua, motivado pelas condições de produção discursiva. Por sua vez, as **condições de produção** referem-se às restrições que a situação pragmática impõem à produção de um enunciado e não outro pelos interlocutores. De certa forma, as pessoas estão sempre em condições de produção discursiva específicas. O fato de que essas mesmas pessoas possuam uma memória discursiva repleta de interdiscursos constantemente em funcionamento nos leva a acreditar que o sentido daquilo que é dito não é alterado somente por mudança do objeto de referência, mas sim porque houve uma mudança na situação discursiva. Essa situação envolve, por sua vez, um domínio, uma posição, um sujeito suposto e um ouvinte instituído e são esses vários fatores o que constitui a chamada condições de produção discursiva.

Em resumo, podemos dizer que quando se trata de abordar o conhecimento pragmático-discursivo dos elementos lingüísticos, deve-se procurar responder a questões do tipo:

quem são os sujeitos envolvidos na situação discursiva? O que querem dizer esses sujeitos? Qual é o contexto da enunciação? Nesse caso, estamos diante dos elementos que compõem o material pragmático dos enunciados: o contexto e a intenção. Adicionalmente, podemos formular as questões: com que autoridade esse discurso foi produzido? Que elementos ideológicos podem ser depreendidos do discurso? Por que este enunciado está aqui e não outro? As informações recuperadas certamente apontarão para o efeito de sentido de um enunciado, levando-se em conta a posição, a situação, a condição e a formação discursiva que é material da Análise do Discurso.

É importante salientar, nesse momento, que esse tipo de abordagem da Análise do Discurso que apresentamos acima trabalha com questões bastante além das preocupações do processamento automático das línguas, haja vista o fato de sublinharem os efeitos que a ideologia provoca na produção e interpretação do discurso. Nesse sentido, grande parte do que o PLN entende por informações da ordem do discurso são, na verdade, noções presentes no campo da Lingüística Textual – uma área de investigação que privilegia o texto e não o discurso como o seu enfoque científico. Dentre os elementos do texto tratados pela Lingüística Textual, um deles é especialmente caro à Lingüística computacional: **os marcadores discursivos** e as suas relações com a *coesão* e *coerência* textual. Com esse princípio procura-se identificar na unidade lingüística algumas marcas formais, como os conectivos, que tornam o texto uma construção coesa – isto é, com unidade de sentido – e coerente – ou seja, sem interferência de ruídos como a contradição.

A aplicação desse conhecimento em PLN é fundamental especialmente quando a ênfase do processamento são as referências anafóricas que chamam a língua para os sentidos já desenvolvidos no decorrer do discurso (ex.: ele, isso, etc.) e os dêiticos que chamam a língua para o contexto, para as circunstâncias do enunciado (ex.: hoje, daqui a pouco, lá, etc.).

4. A Arquitetura de Sistemas de PLN

A arquitetura de um sistema computacional que processa língua natural pode variar de acordo com as especificidades da aplicação. Um exemplo de aplicação para a qual um sistema é o mais completo (e complexo!) possível é o de um tradutor automático. Vamos supor um sistema que traduz uma sentença escrita em português para uma sentença escrita em inglês. Do ponto de vista de suas funções, esse tipo de sistema terá que ser capaz de:

(a) Reconhecer (extrair) cada uma das palavras da sentença em português;

(b) Analisar sintaticamente a sentença, ou seja, associar a cada palavra seus atributos e funções sintáticas;

(c) Representar a sentença numa forma intermediária que agrega as informações levantadas anteriormente;

(d) Analisar semanticamente a sentença, ou seja, extrair um significado global da mesma, a partir dos significados das palavras ou grupos de palavras, e das relações entre elas;

(e) Mapear (associar) o significado extraído em uma representação adequada. Essa representação pode ser independente da língua destino (uma interlíngua, por exemplo), ou não. No caso negativo, pode haver uma transferência da estrutura obtida em (d) para uma estrutura equivalente, de acordo com regras dependentes da língua destino ("tradução por transferência"), ou ainda um mapeamento direto, de palavras ou grupos de palavras da língua origem para seus equivalentes na língua destino ("tradução por método direto"), sendo que nesse caso não há uma representação intermediária.

(f) Transformar a representação anterior em uma sentença na língua destino.

Reconhecemos, no processo acima, duas fases que, mesmo ocorrendo isoladamente, já são bastante complexas: **a fase de Interpretação**, na transformação da sentença na língua origem em uma forma intermediária (passos a, b, c, d), **e a fase de Geração** (e, f) da sentença na língua destino a partir da forma intermediária da sentença original.

Vários aplicativos de PLN possuem apenas uma dessas fases. Por exemplo, em um sistema de consulta a bases de dados, a interface de comunicação com o usuário pode

apresentar somente o módulo de interpretação. Neste caso, o sistema interpreta as perguntas do usuário, obtendo uma representação interna que permita o acesso à base de dados. Uma vez obtidas as respectivas respostas, o sistema as apresenta diretamente ao usuário, sem proceder a qualquer processamento de geração automática para colocar tais respostas em algum formato especial de apresentação ao usuário. Essa mesma interface poderia, por outro lado, apresentar tanto o módulo de interpretação quanto o de geração textual, sendo que este seria responsável por transformar os dados obtidos na consulta em texto (em geral, em interfaces desse tipo a produção textual se resume a utilizar textos pré-fixados, esquemáticos, chamados de *canned texts*).

Sistemas que possuem somente a fase de geração textual, em geral, são aqueles cujas informações não podem ser diretamente apresentadas aos usuários por estarem em um formato ilegível ou de difícil compreensão. Este é o caso de sistemas especialistas: em geral, as respostas a perguntas de usuários de tais sistemas são obtidas utilizando-se os mesmos métodos de obtenção das conclusões dos sistemas (p.ex., o raciocínio lógico-dedutivo) e, por essa razão, nem sempre se encontram em um formato legível para o usuário. Pode-se, então, acoplar uma interface em linguagem natural cuja principal função é gerar um texto que espelhe as informações em formato interno, obtidas a partir das perguntas dos usuários.

Outros sistemas computacionais não chegam a apresentar qualquer uma das fases de interpretação ou geração completamente. Na verdade, eles necessitam de apenas alguns dos passos descritos acima. É o caso, p.ex., dos revisores gramaticais, que não necessitam compreender a sentença, mas apenas extrair sua estrutura sintática. Mesmo neste caso, em algumas situações o conhecimento sobre o significado de palavras se faz necessário.

Os passos do processo de tradução acima ilustram, portanto, as fases de um sistema de Interpretação e Geração de língua. De modo geral, a interpretação é altamente dependente das características dos textos de origem, que incluem não só suas características de superfície, tais como escolhas léxicas ou sintáticas, ou escolhas da ordem dos componentes de uma sentença, como também as características subjacentes à forma textual, expressas por meio das escolhas superficiais pelo produtor do texto com a finalidade de atingir seu objetivo comunicativo. O processo de interpretação deve,

portanto, recuperar, por meio das características de superfície, não só o conteúdo informacional do texto, como também o teor da mensagem, i.e., seu caráter comunicativo. Por essa razão, o processo de interpretação deve ser sofisticado o suficiente para produzir uma representação que seja o mais fiel possível ao texto original, resultando, em geral, em um processo cuja complexidade é proporcional às variações sintáticas, semânticas e pragmáticas permitidas na língua. Assim, o esforço para se obter uma interpretação possível é bastante grande, podendo mesmo haver mais de uma interpretação aceitável (em casos de ambigüidade, por exemplo).

Considere agora uma representação formal (que pode ser bastante complexa) que represente o conteúdo informacional de uma ou mais sentenças a serem produzidas pelo computador. Considere também que a tarefa de geração automática consiste em expressar - ou "realizar" - esse conteúdo na forma textual, lançando mão das possíveis maneiras disponíveis na língua em uso, expressas por sua gramática. Dependendo do objetivo de comunicação, freqüentemente podemos decidir por um conjunto finito e, muitas vezes, simples, de alternativas de regras gramaticais para expressar o conteúdo informacional. Eventualmente, se a aplicação permitir, podemos adotar apenas um padrão sintático. Por exemplo, se a aplicação envolver apenas proposições declarativas, o sistema pode produzir apenas sentenças na voz ativa ou, ao contrário, na voz passiva, e assim por diante, podemos escolher o padrão de cada um desses componentes.

Em outras palavras, podemos, eventualmente, delimitar as opções de geração dentre as inúmeras maneiras de se expressar (escrever) um certo conteúdo informacional em uma dada língua, fazendo com que a tarefa de geração automática seja controlável pelo sistema. Neste caso, o processo se torna dependente das especificações de entrada que, dadas claramente, fazem com que a tarefa de geração seja mais simples do que a de interpretação. Este é o caso, p.ex., de sistemas que têm a função exclusiva de transmitir informações constantes em uma base de dados (sistema de consultas a bases de dados com interface em língua natural, como já exemplificado anteriormente), ou seja, de sistemas cuja função comunicativa principal seja a declarativa ou informativa. Entretanto, para outras aplicações, a delimitação do poder de geração visando a simplificação do sistema pode prejudicar o resultado, dado que a geração textual não implica somente a manipulação do conteúdo informacional, mas também a manipulação dos aspectos comunicativos, segundo as intenções do produtor do texto. Este é o caso,

p.ex., da tradução automática, que exige a correspondência mais fiel possível entre o texto de origem e o texto de destino. Para casos dessa natureza, temos, portanto, um grau de complexidade igualável, se compararmos um sistema de geração com um sistema de interpretação.

Vejamos a seguir as arquiteturas dos sistemas de interpretação e geração de língua natural.

4.1 Arquitetura de um Sistema de Interpretação de Língua Natural

A arquitetura geral de um sistema de interpretação de língua natural é dada pela Figura 4.1. Os módulos de processamento aparecem delimitados por retângulos, enquanto que o conhecimento específico, i.e., os recursos necessários ao processamento, de ordem lingüística (gramática, léxico) ou não (modelos do domínio e do usuário), aparecem delimitados por elipses. Tais recursos são também necessários durante a fase de geração, como veremos na subseção 4.2.

Vale ressaltar que aplicações que não requerem a interpretação de uma sentença têm sua arquitetura simplificada, eliminando-se alguns dos módulos e/ou bases de conhecimento que aparecem na Figura 4.1. Além disso, diferentes aplicações podem exigir algum processamento adicional, que não figura nessa arquitetura. Um exemplo é o processamento morfológico, sobre o qual comentaremos logo a seguir.

Não iremos, aqui, detalhar todos os formalismos que podem ser usados em cada um dos módulos ou fases da interpretação, mas vamos ilustrar, através de exemplos, as funções e a complexidade de cada um dos componentes dessa arquitetura. Antes, no entanto, vamos resumir a função de cada processo presente na arquitetura apresentada. Os recursos lingüísticos necessários para sua operação são descritos na subseção 4.3.

- **Analisador Léxico (ou *Scanner*):** Este processo envolve a identificação e separação dos componentes significativos da sentença sob análise, comumente chamadas de *tokens*, tais como as palavras e os símbolos de pontuação, assim como a associação de atributos ou traços gramaticais e/ou semânticos a cada *token*, com base em consultas ao Léxico. Ele pode ser bastante simples, dependendo da estrutura do léxico e dos atributos requeridos pela aplicação. Pode ser necessária, p.ex., uma

etapa de processamento morfológico anterior ou concomitante com a análise léxica, para a extração de atributos a partir da morfologia dos componentes sentenciais. Isso acontece, p.ex., quando o léxico é composto apenas por formas analisadas da língua (e, portanto, quando o componente sentencial precisa sofrer uma modificação antes de ser associado ao seu verbete) ou quando o léxico é híbrido, contendo formas analisadas e não analisadas[6].

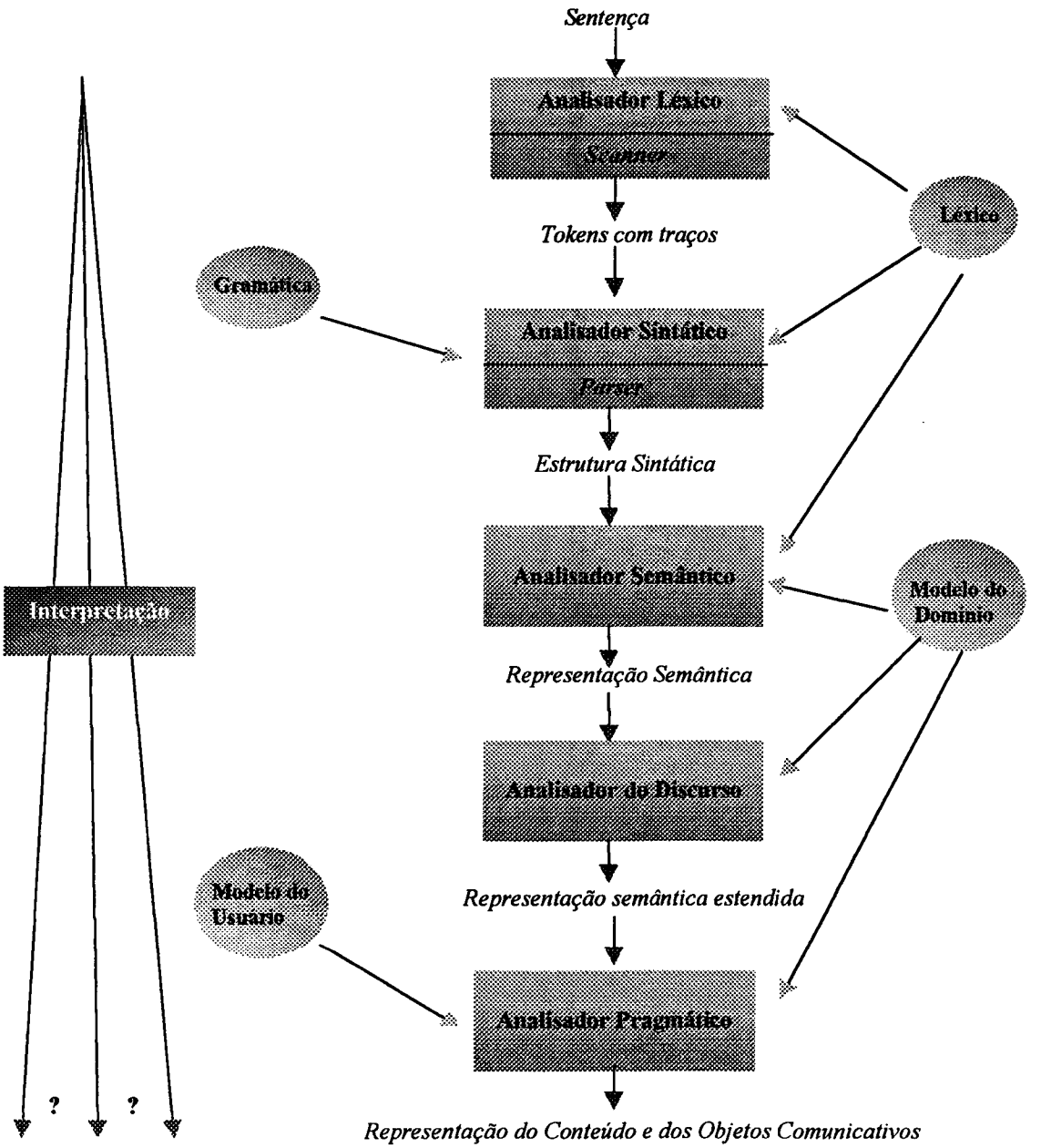


Figura 4.1. Arquitetura de um Sistema de Interpretação de Língua Natural

[6] Formas analisadas são aquelas correspondentes aos verbetes da língua, comumente apresentados em um dicionário; formas não analisadas são aquelas que mantêm uma correspondência fiel à sua forma

- **Analisador Sintático (ou *Parser*):** Este processo é responsável por construir (ou recuperar) uma estrutura sintática válida para a sentença de entrada, também chamada de estrutura profunda. Para tanto, é guiado por uma representação da gramática da língua em questão. Em se tratando de uma língua natural, em geral adota-se uma gramática "parcial" da língua natural, que, embora não abranja todas as construções da língua, contempla aquelas construções válidas de interesse para a aplicação. Assim, evita-se o grande volume de informações gramaticais que pode aumentar demasiadamente a complexidade de sua representação, bem como complexidade do próprio processo de análise. Várias são as técnicas de *parsing* utilizadas em PLN (veja mais detalhes no Capítulo 5). De modo geral, formalismos mais simples são mais eficientes, porém menos abrangentes. Formalismos mais completos e abrangentes tendem a ser complexos e pouco eficientes. A representação da estrutura sintática gerada pelo *parser* varia de acordo com o formalismo e a gramática adotada. Para efeito de ilustração, no entanto, vamos adotar uma linguagem gráfica de representação da estrutura profunda da sentença sob análise, conhecida por **árvore sintática**.

- **Analisador Semântico:** Este processo é responsável pela interpretação de componentes da sentença ou da sentença como um todo e está presente sempre que a aplicação exigir algum tipo de interpretação. Nesse caso, é necessário conhecimento mais específico do domínio, presente no Modelo do Domínio, p.ex., para distinguir a interpretação correta do termo *manga* (se parte de um vestuário ou objeto comestível). Enquanto a estrutura profunda de uma sentença espelha somente a ordem e a caracterização lingüística de seus componentes (i.e., a organização sintática), a estrutura semântica expressa o inter-relacionamento dos componentes sentenciais em nível de significado, podendo ser representada funcionalmente com base nas combinações entre os componentes semânticos expressos pelos componentes sentenciais, na superfície textual. Por exemplo, para a sentença *João comeu a manga*, podemos ter por estruturas profunda e semântica, respectivamente, as seguintes representações simplificadas:

---

sentencial, no contexto de uso. Por exemplo, o verbo no infinitivo *ver* e sua forma não analisada *viu*.

s(sn(substpr(*João*)),sv(vtd(*comer,passado,3ps*),sn(det(*o*),subst(*manga*))
ação(*comer*,agente(anim(*João*)),objeto(comest(*manga*)))[7]

Vale notar que, para a sentença *João costurou a manga*, a estrutura profunda será similar à estrutura profunda exemplificada acima, com exceção dos valores terminais *comer* e *costurar*. Entretanto, a estrutura semântica será fundamentalmente distinta, já que agora o objeto deixa de ser comestível. Dessa forma, os diferentes significados de sentenças gramaticalmente similares (cujas estruturas profundas são as mesmas, com exceção dos símbolos do vocabulário) são necessariamente expressos em cada estrutura semântica, sendo este o componente principal para a distinção interpretativa. Formalismos de representação semântica em geral diferem dos formalismos gramaticais de *parsing*, sendo que várias linguagens de representação são possíveis[8]. Uma das mais utilizadas é a Lógica de Predicados (Clocksin and Mellish, 1981; Colmerauer, 1977; Kowalski, 1974), adotada no exemplo acima.

- **Analisador do Discurso:** Embora qualquer discurso possa ser mono ou multi-sentencial, para efeito de ilustração estamos considerando aqui somente os do último tipo para discutir o problema da análise discursiva. Neste caso, o significado de uma sentença pode depender das sentenças que a antecedem e pode influenciar os significados das sentenças que a seguem. Em geral, em textos multi-sentenciais são utilizados recursos lingüísticos que tornam a resolução analítica mais complexa. Por exemplo, para fazer o texto "fluir" ou tornar-se estilisticamente mais elegante, é comum utilizarem-se referências anafóricas (p.ex., por meio de pronomes: *ele, ela, este, aquela*, ou por meio de sinônimos: *a menina*, referindo-se a *Amélia*), referências dêiticas, cujos componentes indicados são extratextuais (p.ex., *aqui, ali, hoje*) ou outras figuras de discurso. O analisador de discurso trata exatamente desse tipo de inter-relacionamento, assumindo maior importância à medida que aumenta a complexidade de resolução das associações entre os componentes sentenciais. Para a resolução, p.ex., de referências pronominais ou dêiticas, o analisador pode utilizar as

[7] Leia as abreviações como: s - sentença; sn - sintagma nominal; substpr - substantivo próprio; sv - sintagma verbal; vtd - verto transitivo direto; 3ps - 3a. pessoa do singular; det - determinante; subst - substantivo; anim - objeto animado (ou ser humano, no caso); comest - objeto inanimado comestível.

[8] Para saber mais sobre diferentes formalismos de representação semântica, veja (Rich and Knight, 1993; Shieber, 1986; Winston, 1993; Woods, 1986). Modelos semânticos, em geral, podem ser encontrados em (Grosz et al., 1986).

noções de *foco do discurso*, que deve ser reconhecido com base em preferências sintáticas ou semânticas. Repare as marcas dos focos nas diferentes construções para uma mesma proposição-pergunta: *Foi José quem pegou o livro?* e *Foi o livro o que José pegou?*. O analisador de discurso, em geral, estende a representação semântica produzida pelo analisador semântico com as anotações sobre as figuras de discurso.

- **Analisador Pragmático:** Apesar de vários níveis de análise de uma estrutura superficial de um texto permitirem a obtenção de uma representação do significado (representação semântica, conforme ilustrada na Figura 4.1), a obtenção da mensagem original, como resultado da interpretação, propriamente dita, pode ainda estar sujeita a aspectos pragmáticos da comunicação. Por exemplo, nem sempre o caráter interrogativo de uma sentença expressa exatamente o caráter de solicitação de uma resposta. Suponha que a sentença *"Você sabe que horas são?"* possa ser interpretada como uma solicitação para que as horas sejam informadas ou como uma repreensão por um atraso ocorrido. No primeiro caso, a pergunta informa ao ouvinte que o falante deseja obter uma informação e, portanto, expressa exatamente o caráter interrogativo. Entretanto, no segundo caso, o falante utiliza o artifício interrogativo como forma de impor sua autoridade. Diferenças de interpretação desse tipo claramente implicam interpretações distintas e, portanto, problemáticas, se não for considerado o contexto de ocorrência do discurso.

Os limites entre os cinco processos anteriores (léxico, sintático, semântico, discursivo e pragmático) são normalmente obscuros. Esses processos nem sempre são executados seqüencialmente, posto que as informações são interdependentes e, logo, podem ser executadas concomitantemente. Considere, p.ex., a sentença *"É o pote creme de molho inglês?"* (exemplo extraído de Rich and Knight, 1993, p.437). Durante sua análise sintática, é preciso decidir qual é o sujeito e qual é o predicado, dentre os três substantivos da sentença (*pote, creme e molho*) e dar a ela o formato *"É x y?"*. Lexicamente, todas as seguintes delimitações da frase *pote creme de molho inglês* são possíveis: o pote, o pote creme, o pote creme de molho, o pote creme de molho inglês, creme de molho inglês, molho inglês, inglês. Entretanto, o processador sintático será incapaz de decidir quais, dentre essas formas, correspondem a estruturas sintáticas válidas, se não contar com algum modelo de mundo em que certas estruturas fazem sentido e outras não. Caso esse modelo exista no sistema automático, é possível obter-se

uma estrutura que permita, p.ex., a interpretação *o pote de cor creme contém molho inglês*, e não *o pote é creme de molho inglês*. Desse modo, as decisões sintáticas dependem da análise do discurso ou do contexto de uso e, portanto, os processos representados na Figura 4.1 interagem entre si. Não é difícil notar que a execução seqüencial dos processos de interpretação simplifica sobremaneira o projeto do sistema, se considerarmos que o resultado de uma fase constitui a entrada para a fase subseqüente. Neste caso, os processos se tornam modulares e, portanto, o controle é menos complexo. As decisões sobre a seqüencialização ou combinação dos processos dependem das características do projeto particular que se tem em mente.

Vamos agora ilustrar o processo de interpretação com a análise da seguinte sentença: *O menino viu o homem de binóculo*. Trata-se de uma sentença ambígua da língua portuguesa, uma vez que pode ser interpretada como se (a) O menino estivesse com o binóculo, ou (b) O homem estivesse com o binóculo. Essa ambigüidade é dita **sintática**, e se dá quando uma mesma sentença pode ser mapeada em mais de uma estrutura sintática válida. Esse tipo de ambigüidade só pode ser tratado por gramáticas que sejam capazes de gerar mais de uma estrutura sintática para a mesma cadeia de entrada. A Figura 4.2 mostra as árvores de derivação sintática para duas das interpretações acima. Outro tipo de ambigüidade possível é a **lexical** (também chamada de semântica), que se dá quando uma palavra pode ser interpretada de mais de uma maneira. Por exemplo, a sentença *João procurou um banco.* pode se referir à procura de um banco financeiro ou de um lugar para se sentar.

Alguns exemplos de entradas do Léxico para esse exemplo são apresentados abaixo. Utilizamos aqui o formalismo PATR-II (Shieber, 1984).

menino .
<categoria> = substantivo
<gênero> = masculino
<número> = singular

viu .
<categoria> = verbo
<tempo> = passado
<número> = singular
<pessoa> = 3
< arg1> = SN|SV[9]

[9] Caso em que o verbo admite também uma forma verbal como objeto direto.

o .

<categoria> = determinante
<gênero> = masculino
<número> = singular

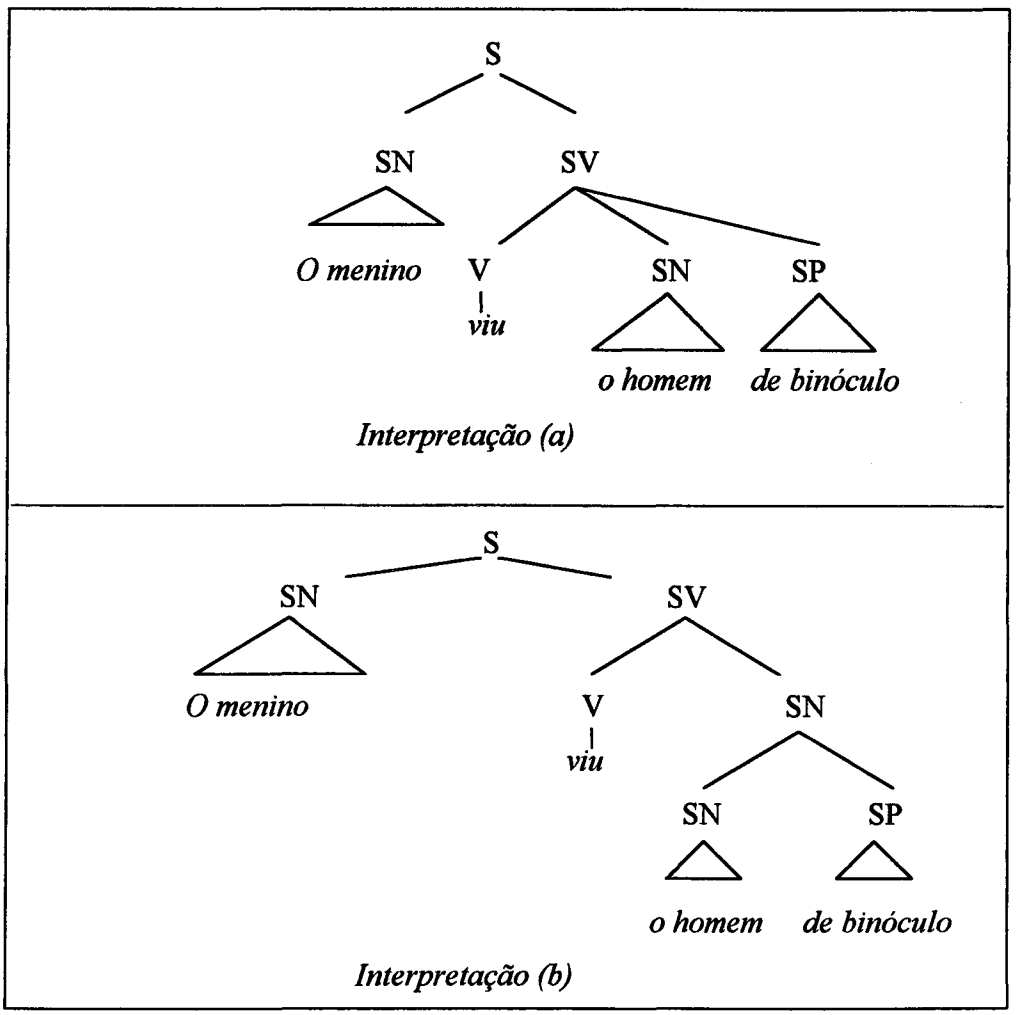


Figura 4.2. Exemplo de Ambigüidade Sintática

Em aplicações para as quais informações semânticas no nível lexical são relevantes, um conjunto de **traços semânticos** poderia ser associado a cada item lexical. Por exemplo:

menino→ [+humano], [+jovem]
homem→ [+humano], [-jovem]
binóculo→ [+inanimado], [+concreto]

Uma parte da Gramática para a análise do exemplo acima é dada pelas seguintes regras de produção:

S → SN SV
SN → Det Subst
SN → SN SP
SV → V SN

SV → V SN SP
SP → Prep Subst

Finalmente, uma possível representação semântica para a sentença de interpretação (a) poderia ser baseada em relações semânticas, p.ex.:

agente(ação(ver), menino)

objeto(ação(ver), homem)

instrumento(ação(ver), binóculo)

Repare ainda que, se essa sentença fosse parte de um texto, p.ex., "*João ganhou um binóculo de seu pai. O menino viu o homem de binóculo.*", o processo de interpretação deveria ser capaz de resolver a referência entre "menino" e "João" e, ainda, determinar que "homem" não se refere nem a "João" nem ao "pai de João". Este tipo de decisão é de responsabilidade do analisador de discurso. O analisador pragmático, nesse exemplo, teria atuado juntamente com a análise sintática, para definir a estrutura sintática mais provável e, assim, eliminar a ambigüidade da sentença.

4.2 Arquitetura Geral de um Sistema de Geração de Língua Natural

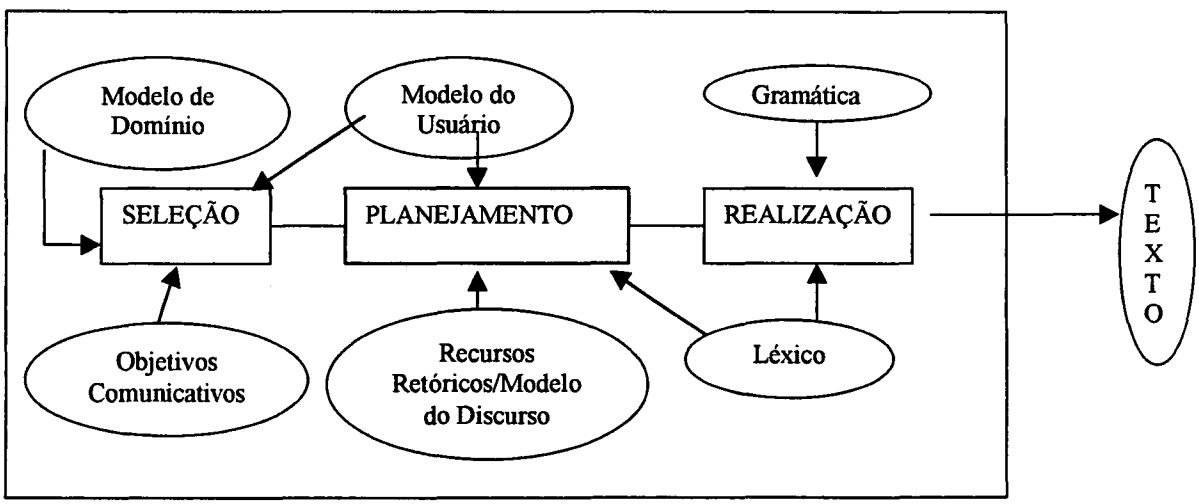


Figura 4.3. Fases Principais de um Gerador de Textos[10]

A arquitetura geral de um sistema automático de geração de língua natural pode ser dada pelo esquema da Figura 4.3, que ilustra um gerador comumente chamado de *gerador de três passos*, por considerar três processos fundamentais durante a geração: a **Seleção** de conteúdo, o **Planejamento** da estrutura textual (ou planejamento do texto) a

[10] Arquitetura extraída de (Matthiessen and Bateman, 1991).

partir do conteúdo selecionado e a **Realização** da estrutura de texto em texto, propriamente dito (ou realização textual). Embora a geração textual lance mão de recursos similares aos da interpretação (como ilustra a Figura 4.3 e também será explicitado na Seção 4.3), já que o PLN envolve o mesmo conhecimento lingüístico ou extralingüístico, independentemente do processo em questão ou do grau de profundidade de seu uso, os mecanismos de processamento da informação são, em geral, distintos o suficiente para não permitirem a inversão dos módulos de um processo ao outro. Assim, na maioria das vezes não é possível considerar a geração como o processo inverso da interpretação (embora uma arquitetura dessa natureza venha sendo investigada há bastante tempo).

Ao contrário de um sistema de interpretação, um gerador tem a função de produzir textos[11] em língua natural a partir de um conjunto de elementos de conteúdo e de objetivos de comunicação. Em muitas aplicações de PLN, no entanto, a geração de língua natural é feita de uma maneira bastante simplificada, em que os textos são construídos pela justaposição de partes (ou segmentos) textuais pré-determinadas (e, neste caso, já definidas durante a fase de projeto do sistema). Outras vezes, esquemas de texto, conhecidos como *canned texts*, são "preenchidos" de forma a compor o texto final. Neste caso, os esquemas são também pré-definidos, mas possuem uma parte variável que somente pode ser determinada em tempo de processamento. Apesar das limitações inerentes a essas técnicas, para muitas aplicações elas se mostram bastante satisfatórias. É o caso, p.ex., de respostas a consultas a bases de dados, que são geralmente simples e, portanto, não exigem um processamento mais sofisticado (e caro!).

As tarefas ilustradas acima envolvem o controle sobre a variedade de formas lingüísticas usadas para expressar o conteúdo selecionado e sobre a organização desse conteúdo, ou estruturação do texto, que são tarefas equivalentes às de um produtor (humano) de textos. Ao contrário da gramática de interpretação, a gramática de geração procede a partir das funções dos elementos conceituais do texto para produzir a estrutura textual e, portanto, seus elementos lingüísticos. Dessa forma, decisões sobre o

[11] Consideramos que um texto pode ser mono ou multi-sentencial. Entretanto, a geração de sentenças isoladas não reflete a complexidade e os desafios da geração multi-sentencial, visto que esta envolve questões de coerência e coesão do discurso que, no caso mono-sentencial, são fortemente simplificadas.

vocabulário, os constituintes sintáticos e a própria forma da sentença são de responsabilidade do gerador a partir do instante em que se determina como combinar os conceitos do discurso a fim de atingir os objetivos de comunicação desejados. Nesse processo, os componentes do domínio e do discurso são utilizados a fim de se acessar o Léxico e a gramática sob enfoque, para determinar os componentes textuais. Em outras palavras, parte-se, em geral, de uma representação profunda do discurso a fim de se obter a representação superficial, conforme veremos a seguir, pela especificação funcional de cada uma das fases ilustradas acima.

- **Seleção do conteúdo:** Este processo tem a função de selecionar os itens de conhecimento que deverão fazer parte do texto. Por exemplo, numa aplicação de geração de respostas em língua natural a consultas a uma base de dados, isso equivale a extrair, do registro selecionado da base, os itens de dados que comporão a resposta (p.ex., o nome, a idade e o RG de um funcionário). O Modelo do Usuário pode determinar a quantidade de informação necessária na resposta. Diz-se, então, que essa fase determina, num processo comunicativo, "o que dizer".

- **Planejamento do texto:** Esse componente do gerador de textos, também chamado de componente estratégico, é responsável por planejar a comunicação. É nela que se decide "quando dizer" o que foi selecionado na fase anterior. A entrada para o planejador pode ser bastante variada e depende da aplicação que contém o gerador. Por exemplo, o conteúdo informacional pode estar representado em forma de registros de uma base de dados, de uma tabela de registros, de proposições lógicas, etc. Nessa fase, o conteúdo deve ser organizado para uma melhor apresentação textual. Isso implica na adição de especificações retóricas, na determinação da seqüência em que as informações serão apresentadas e em algumas decisões sobre a escolha de palavras a serem usadas. O planejador produz uma forma intermediária do texto, chamada de **plano do texto**. Os formalismos de representação do plano do texto diferem entre si quanto ao aspecto do plano que privilegiam. Alguns privilegiam a estrutura retórica do texto, como por exemplo, a Rhetorical Structure Theory, RST (Mann and Thompson, 1986). Outros privilegiam os aspectos pragmáticos, como o gerador Pauline (Hovy, 1988).

- **Realização do texto:** Esse processo, também chamado de componente tático, componente lingüístico ou gerador de superfície, é responsável pela realização gramatical do plano de texto produzido pelo planejador. Nesta fase, podemos ter dois subprocessos distintos: a *determinação* dos itens lingüísticos, propriamente dita, e sua *linearização*, i.e., a "planificação" da estrutura textual pela seqüencialização de tais itens na forma textual, produzindo um encadeamento de sentenças válidas na língua em foco. As contribuições desse componente para o processo de geração envolvem as seguintes decisões lingüísticas e decisões sobre o conhecimento do domínio e do discurso:

- Escolha de vocabulário;
- Escolha do estilo do texto (p.ex., prosa, diálogo, etc.);
- Escolhas léxicas, morfológicas e sintáticas adequadas para expressar conteúdo e estrutura textual;
- Escolha de figuras de discurso para manifestar apropriadamente as intenções do falante/escritor (questão de foco, ênfase, etc.);
- Escolhas que garantam a coesão do discurso, i.e., a fluidez do texto (p.ex., o uso de marcadores de seqüencialização das informações);
- Escolhas que garantam a coerência do discurso, i.e., que expressem o inter-relacionamento retórico/semântico desejado (p.ex., o uso de uma marca de contraste entre componentes textuais que devem ser contrastados);
- E, finalmente, decisões de linearização, p.ex., ordenação das informações, concordância gramatical, etc.

Como podemos ver, além da fase de lexicalização (escolha de palavras), o realizador possui como funções: (a) mapear a estrutura temática de cada sentença em uma estrutura sintática de superfície; (b) aplicar regras gramaticais, como a concordância entre sujeito e verbo, entre determinante e substantivo; (c) escolher as palavras das classes fechadas (pronomes, conjunções, artigos, etc.); (d) flexionar as palavras de classes abertas, como a conjugação verbal; e (e) linearizar a árvore sintática em uma cadeia de palavras flexionadas. Se considerarmos, p.ex., um plano de texto similar à estrutura profunda acrescida de informações de discurso, já ilustrada na subseção

anterior, a fase de linearização corresponderá ao percurso da árvore em profundidade-primeiro, da esquerda para direita.

Normalmente, a distinção entre as fases de planejamento e realização é vantajosa porque provê pelo menos dois níveis de abstração, de modo que detalhes que são relevantes para o realizador podem ser ignorados pelo planejador. Por exemplo, a decisão sobre qual determinante usar no contexto de uma proposição não é uma consideração apropriada no momento em que se escolhe uma estratégia para convencer o leitor dessa proposição. Mais ainda, se a interface entre esses dois níveis for cuidadosamente especificada, parece possível construir um componente estratégico geral que possa ser usado para uma grande variedade de aplicações, mesmo que o componente tático seja variável (p.ex., quando se deseja obter um sistema de geração multilingual).

Da mesma forma que na interpretação, podemos considerar três modos distintos de interação entre os processos de um gerador automático: (a) o **seqüencial** (ou *geração em pipeline*), em que os três processos ilustrados são estritamente seqüenciais (p.ex., a realização acontece apenas quando o planejamento já terminou) e, portanto, a atuação de cada módulo não interfere na do outro; (b) o **intercalado** (ou *interleaved generation*), em que os módulos executam suas funções de modo intercalado, intercomunicando-se entre si à medida que cada processo necessita tomar decisões que envolvem outras esferas de conhecimento (p.ex., decisões sintáticas dependentes do conhecimento do usuário) - neste caso, a intercomunicação ocorre por demanda, i.e., somente quando um módulo acusa a necessidade de outras informações que não são de sua responsabilidade e (c) o **combinado** (ou *merged generation*), em que os processos executam todas as tarefas sem que seja possível distinguir ou modularizá-las.

Veja mais detalhes sobre geração de texto em (Appelt, 1985; Dale, 1992; Paris et al., 1991; MacDonald and Bolc, 1988; McKeown, 1985; McKeown and Swartout, 1987; Dale et al., 1990; Smadja and McKeown, 1991 e Matthiessen et al., 1991).

4.3. Recursos lingüísticos para o processamento de línguas naturais

Os recursos lingüísticos presentes nas arquiteturas de interpretação e geração são detalhados a seguir.

- **Léxico:** Consiste em um conjunto de palavras ou expressões da língua associadas a um conjunto de atributos, ou traços morfossintáticos, e traços semânticos (opcionais). Durante a interpretação, o léxico é acessado pelos analisadores léxico, sintático e semântico (vide Figura 4.1), cada um deles visando funções específicas, sendo que as suas principais tarefas são, respectivamente: reconhecer as *tokens* da sentença de entrada e recuperar seus principais traços (p.ex., *comida* → *token* = *comer, categoria=verbo/substantivo, gênero=fem, número=sing*); reconhecer ou atribuir categorias sintáticas às *tokens*, para a obtenção da estrutura profunda da sentença (p.ex., *comida* → *token* = *comer, categoria=verbo, tempo=particípio passado, gênero=fem, número=sing*) e verificar a validade do relacionamento semântico da *token* sob análise em função do contexto em que ela ocorre na sentença, i.e., em relação às demais *tokens* obtidas durante a análise dos demais componentes sentenciais. Neste caso, o léxico deve fornecer, além dos traços gramaticais, os traços semânticos de suas entradas, para possibilitar a verificação semântica. Seu tamanho ou número de entradas lexicais e a estrutura de suas entradas podem variar de acordo com a natureza da aplicação. Existem vários formalismos de representação da informação que constitui o léxico, porém, é necessário que a representação adotada esteja de acordo com o formalismo escolhido para a representação da gramática, ou possa ser compreendido pelo processo de manipulação da mesma, uma vez que ambos os processos - de acesso e manipulação do léxico e de manipulação das regras gramaticais - interagem entre si, tanto na interpretação quanto na geração.

- **Gramática:** Em geral representada por um conjunto de regras gramaticais, a gramática define quais são as cadeias de palavras válidas (i.e., sentenças) em uma língua natural. Há vários tipos de gramáticas e diversos formalismos de representação computacional[12]. Quase todos, no entanto, podem ser expressos por regras de produção, do tipo S → SN SV. A leitura de regras de produção desse tipo pode ser realizada em função do tipo de manipulação que se pretende. Por exemplo, quando essa regra for usada durante o processo de interpretação, ela pode ser

[12] Veja sobre os diferentes tipos de gramáticas e formalismos de representação gramatical em (Rich and Knight, 1993; Shieber et al., 1986; Winston, 1993; Woods, 1986). Modelos sintáticos, em geral, podem ser encontrados ainda em (Grosz et al., 1986).

entendida como *Para reconhecer uma sentença S reconheça como seus componentes um sintagma nominal, SN, seguido por um sintagma verbal, SV.* Se a mesma regra for utilizada em um processo de geração, ela pode ser lida como *Uma sentença S pode ser constituída por um sintagma nominal, SN, seguido de um sintagma verbal, SV.* Desse modo, a partir de uma única especificação da gramática em uso, pode-se proceder a uma aplicação específica, quer seja ela de interpretação ou de geração. Entretanto, vale notar que nem sempre um mesmo formalismo de representação das regras gramaticais da LN dará origem a um único mecanismo computacional que valha tanto para a interpretação quanto para a geração, uma vez que a computação em um e outro caso nem sempre é intercambiável.

- **Modelo do Domínio:** Este módulo fornece conhecimento sobre o domínio específico da aplicação, p.ex., informações de senso comum sobre as entidades do discurso em foco (como *o homem é mortal*, ou *animado(homem)*), padrões ontológicos sobre o modelo do domínio (como uma taxonomia do mundo animal), etc. Essas informações servirão tanto à interpretação quanto à geração. No primeiro caso, fornecendo subsídios para o correto inter-relacionamento semântico entre os componentes sentenciais, para a desambigüização lexical (p.ex., para a desambigüização de *manga* como parte de um vestuário ou objeto comestível, como já exemplificamos antes) ou para a determinação de figuras de estilo ou figuras retóricas particulares, durante a análise do discurso. Diversas linguagens de representação do conhecimento podem ser utilizadas neste módulo, dentre as quais destacamos a lógica de predicados, as redes semânticas (Quillian, 1968; Rumelhart and Norman, 1975; Simmons, 1973; Woods, 1986), os *frames* (Minsky, 1975), entre outras.

- **Modelo do Usuário:** Em sistemas de PLN, o modelo do usuário permite que se configure o contexto de ocorrência do discurso de modo a prever ou reconhecer características que levem a determinações específicas da estrutura ou do significado textual. Por exemplo, o grau de informatividade na geração textual depende do que é relevante ao leitor e, portanto, irá implicar escolhas diversas de vocabulário, estruturas lingüísticas, etc.; o nível de conhecimento do assunto (superficial ou

profundo) que o usuário apresenta pode levar a estruturas semânticas particulares, que, resultantes de um processo de *parsing*, podem auxiliar um sistema de consulta a, p.ex., fornecer respostas em grau adequado de clareza. Em geral, o conhecimento representado nesse módulo inclui as seguintes informações a respeito do usuário do sistema: seus objetivos, planos, preferências, intenções, etc. Linguagens formais de representação de tal conhecimento incluem, p.ex., planos e *scripts* (Schank and Abelson, 1977) ou atos de fala (Grice, 1975).

5. Processamento Sintático

Entre as etapas que caracterizam o processamento automático das línguas naturais, uma é particularmente sintomática das limitações da máquina e da complexidade dos fenômenos lingüísticos. Trata-se do processamento sintático, que reúne hoje algumas das principais divergências teóricas e metodológicas que cercam a lingüística computacional. O objetivo deste capítulo é passar em revista, de forma bastante esquemática, os principais tópicos relacionados ao processamento sintático automático da língua portuguesa, considerando, primeiramente, as categorias sintáticas que chegam da teoria lingüística, e verificando, em seguida, sua aplicabilidade na prática lingüístico-computacional.

5.1 O que é linguagem?

Para Saussure, por muitos considerado o fundador da Lingüística contemporânea, esta é um pergunta sem resposta. A linguagem seria incognoscível. Estaria perpetuamente dividida em duas faces "que se correspondem e das quais uma não vale senão pela outra" (Saussure, 1988). Seria, a um só tempo, e contraditoriamente, acústica e articulatória, física e psíquica, individual e social, estática e dinâmica:

> *"Tomada em seu todo, a linguagem é multiforme e heteróclita; a cavaleiro de diferentes domínios, ao mesmo tempo física, fisiológica e psíquica, ela pertence além disso ao domínio individual e ao domínio social; não se deixa classificar em nenhuma categoria dos fatos humanos, pois não se sabe como inferir sua unidade." (Saussure, 1988; p.17)*

O objeto de estudo da Lingüística não seria, pois, a linguagem, mas o seu produto social: a língua (ou cada uma das línguas naturais). A **língua** seria a face contratual, autônoma, homogênea e concreta da linguagem. Contratual, porque pressuporia um acordo prévio dos falantes sobre o vocabulário, suas regras de combinação e de uso; autônoma, porque seria auto-consistente, sem a necessidade de referência a outros sistemas semiológicos; homogênea, porque as relações internas à língua seriam estáveis e não poderiam ser modificadas ao sabor dos desejos de cada falante; concreta, porque os signos lingüísticos estariam materializados na fala. Diferentemente do que acontece

em relação à linguagem, se poderia dizer que a língua (ou cada uma das línguas) constitui uma unidade, delimitável, abordável, cujo princípio de unificação, a força centrípeta que mantém a língua unitária, seria o objeto de estudo da Lingüística.

Para Saussure, este princípio de unificação seria o conjunto de relações sintagmáticas e associativas que se estabelecem, em cada língua, entre os signos lingüísticos. **Relações sintagmáticas** seriam aquelas "baseadas no caráter linear da língua, que exclui a possibilidade de pronunciar dois elementos ao mesmo tempo"[13]. São relações que se estabelecem *in praesentia*, como as que operam entre os fonemas /f/ e /a/ em /fala/, entre os morfemas {cant}, {a}, {re} e {mos} em *cantaremos*, ou entre as palavras *Maria* e *morreu* na sentença *Maria morreu*. **Relações associativas**, também chamadas paradigmáticas, são aquelas que se estabelecem na memória do falante, e fazem parte do "tesouro interior que constitui a língua de cada indivíduo"[14]. São relações que unem termos *in absentia*, como aquelas verificáveis entre os fonemas /b/, /c/, /f/ e /g/ no contexto /_ala/, entre os morfemas { }, {re}, {va}, {sse} no contexto {canta_mos}, ou entre as formas *morreu*, *saiu*, *matou Pedro* e *gosta de ir ao cinema*, no contexto *Maria* ____ . Descrever uma língua, segundo Saussure, seria estabelecer o traçado dessas relações sintagmáticas e associativas.

No entanto, como os próprios exemplos acima assinalados o indicam, as relações sintagmáticas e associativas na língua se estabelecem diferentemente para diferentes dimensões do signo lingüístico. Há relações sintagmáticas e associativas entre fonemas, entre morfemas, entre palavras, entre sentenças, entre textos. E as relações são específicas aos objetos lingüísticos relacionados. Dificilmente as relações sintagmáticas entre fonemas terão alguma utilidade na consideração das relações que se estabelecem, por exemplo, entre os morfemas da língua. Da mesma forma, as relações associativas que as sentenças estabelecem na memória do falante não são pertinentes na consideração daquelas que se colocam a partir de contextos morfológicos específicos.

A percepção da diferente natureza das relações sintagmáticas e associativas deu origem a uma fragmentação do sistema lingüístico hoje onipresente. Reconhece-se que as línguas possuem diferentes níveis de significância, aos quais correspondem signos

[13] (Saussure, 1988; p.142)
[14] (Saussure, 1988; p.143)

lingüísticos de diferentes dimensões. Fonemas, morfemas, palavras, sentenças e textos, ainda que pertencentes a um mesmo sistema lingüístico (a língua portuguesa, por exemplo), constituem diferentes **níveis de descrição lingüística**, que conservam autonomia conceitual e metodologia própria, não necessariamente aplicável aos outros níveis do mesmo sistema. Essa concepção estratificada da língua, dividida em seus diferentes níveis de descrição, será o ponto de partida do desenvolvimento de uma série de disciplinas lingüísticas, dedicadas ao estudo específico de cada uma dessas camadas. Assim, a Fonologia se ocuparia do nível do fonema; a Morfologia, do nível do morfema; a Sintaxe, do nível da frase; e a Lingüística do Texto, do nível do texto. Nesta seção estaremos particularmente envolvidos com a recuperação dos desdobramentos do nível sintático e suas implicações para a lingüística computacional.

5.2 A Sintaxe

A partir do que se disse no item anterior, pode-se definir Sintaxe como **a disciplina que estuda as relações sintagmáticas e associativas que se estabelecem nas sentenças de uma determinada língua.** Trata-se, portanto, de investigar quais são as relações que as palavras estabelecem entre si em uma determinada frase e quais são as relações (mnemônicas) que se estabelecem entre palavras em um mesmo contexto. Neste último caso, seremos levados à identificação de um repertório de **categorias lexicais**, também chamadas "partes do discurso". No primeiro caso, chegaremos a um conjunto de **categorias funcionais**, também chamadas "funções sintáticas".

O conjunto das categorias lexicais, que correspondem às relações associativas, pode ser aduzido a partir do estabelecimento de contextos de ocorrência. Por exemplo: a relação associativa que se estabelece entre as formas *Maria, a menina, ela, alguém, o pobre* e todas as outras que podem preencher a posição vaga no contexto "____ *apareceu.*" recebe comumente o nome de "substantivo"[15]. A definição de substantivo seria, portanto, antes negativa, relacional, estrutural. Da mesma forma, se chegará aos conceitos de adjetivo, verbo, advérbio, preposição, conjunção, numeral, pronome, artigo e interjeição, que esgotariam as possibilidades de variação lexical do sistema lingüístico.

[15] A relação será desdobrada em novas subespecificações dos tipos de substantivo a partir do estabelecimento de outros contextos de ocorrência: substantivo próprio, caso de *Maria*; substantivo comum, caso de *menina*; pronome substantivo, caso de *ela* e *alguém*; e adjetivo substantivado, caso de *pobre*.

O conjunto das categorias funcionais remete às relações sintagmáticas. Já não se trata de pensar em relações entre termos ausentes mas entre as unidades consecutivas de uma mesma sentença. Assim, em *A menina deu o livro para o menino* há uma relação (sintagmática) que se estabelece entre *a* e *menina* a que normalmente se dá o nome de "sintagma nominal" (SN, em inglês: NP – *Noun Phrase*). A mesma relação se verifica entre *o* e *livro* e entre *o* e *menino*, que constituem outros sintagmas nominais. Uma relação diferente se estabelece entre a preposição *para* e o sintagma nominal *o menino*. Essa relação recebe o nome de "sintagma preposicional" (SP, em inglês PP – *Prepositional Phrase*). Na mesma sentença, se pode perceber ainda uma relação entre o verbo *deu* e o sintagma nominal e o sintagma preposicional que o seguem: trata-se de um "sintagma verbal" (SV, em inglês VP – *Verbal Phrase*). A gramática tradicional geralmente atribui conteúdos a essas relações, reclassificando-as, de acordo com a posição e com o papel desempenhado na sentença, em "sujeito", "predicado", "objeto direto", "objeto indireto", "adjunto adnominal" e outros.

Categorias lexicais e categorias funcionais são úteis porque permitem descrever e prever uma série de fenômenos sintáticos, como a colocação, a concordância, a regência, a elipse, a topicalização e a apassivação, que são cruciais no reconhecimento e na geração de sentenças da língua portuguesa. O conjunto das previsões desses fenômenos sintáticos constitui a **gramática** de uma língua.

5.3 Formalismos gramaticais

O termo "gramática" deriva do grego γραμμα, que tinha originalmente a acepção de letra, símbolo gráfico que representa os sons da língua. Em pouco tempo, o estudo da gramática passou a contemplar técnicas de escrita e do bem dizer, constituindo um conjunto de regras de bom uso das formas da língua, apoiadas em critérios ora lógicos, ora geográficos, ora literários, ora históricos, ora sociais. É este o sentido normalmente contemplado pelo senso-comum, que entende a gramática como conjunto de regras de boa formação das palavras e das sentenças da língua. No entanto, é importante perceber que a interpretação do termo "boa formação" tem sido bastante discrepante entre lingüistas e gramáticos. Para os últimos, cujo objetivo é predominantemente normativo (de onde a expressão "**gramática normativa**"), estão bem formadas as construções que encontram amparo no uso que os autores da literatura brasileira (particularmente os

autores do século passado) fizeram da língua; para os lingüistas, cuja preocupação central é a descrição da língua (de onde "**gramática descritiva**"), estão bem formadas as construções que possam cumprir o objetivo comunicativo da linguagem, pouco importando se essas formas são ou não abonadas por autoridades lingüísticas e literárias. Assim, *O pessoal foram no cinema e Nós vamos se matar,* embora não sejam aceitas pela gramática normativa, são consideradas sentenças da língua portuguesa para a gramática descritiva.

Tanto a gramática normativa quanto a gramática descritiva se constituem a partir da idéia de **regra**, de que o comportamento lingüístico dos falantes é regular, de que pode ser previsto e modulado a partir do estabelecimento de um conjunto finito de instruções que, em última instância, pode ser ensinado, de forma explícita, tanto aos falantes quanto aos não-falantes da língua. Há várias formas de redigir esse conjunto de regras, sendo a mais conhecida aquela que as gramáticas normativas da língua portuguesa assumem: "*em português, o verbo concorda com o sujeito em número e pessoa*", por exemplo. Definições desta natureza, no entanto, são geralmente ambíguas, envolvendo um grau de interpretabilidade que dificilmente pode ser alcançado por falantes não habituados a essas categorias gramaticais [como de fato ocorre entre os alunos do ensino médio e do ensino fundamental]. O mesmo, de forma ainda mais dramática, se verifica para as máquinas.

O processamento automático das línguas naturais tornará imperiosa a redescrição das regras gramaticais segundo critérios formais, em oposição aos critérios nocionais que são privilegiados pela gramática normativa. De nada adianta informar à máquina que "substantivo é o nome com que designamos os seres em geral — pessoas, animais e coisas[16].", se a máquina não estiver aparelhada para identificar o que são os "seres em geral" e para reconhecer qual a utilidade dos nomes na língua. Uma definição computacional efetivamente válida deve levar em consideração antes a forma do que o conteúdo das categorias lexicais e funcionais.

A mais célebre tentativa de formalização da gramática das línguas naturais é, sem dúvida, a que tem início com o lançamento, em 1957, do livro *Syntactic Structures*, de

[16] (Bechara, 1972)

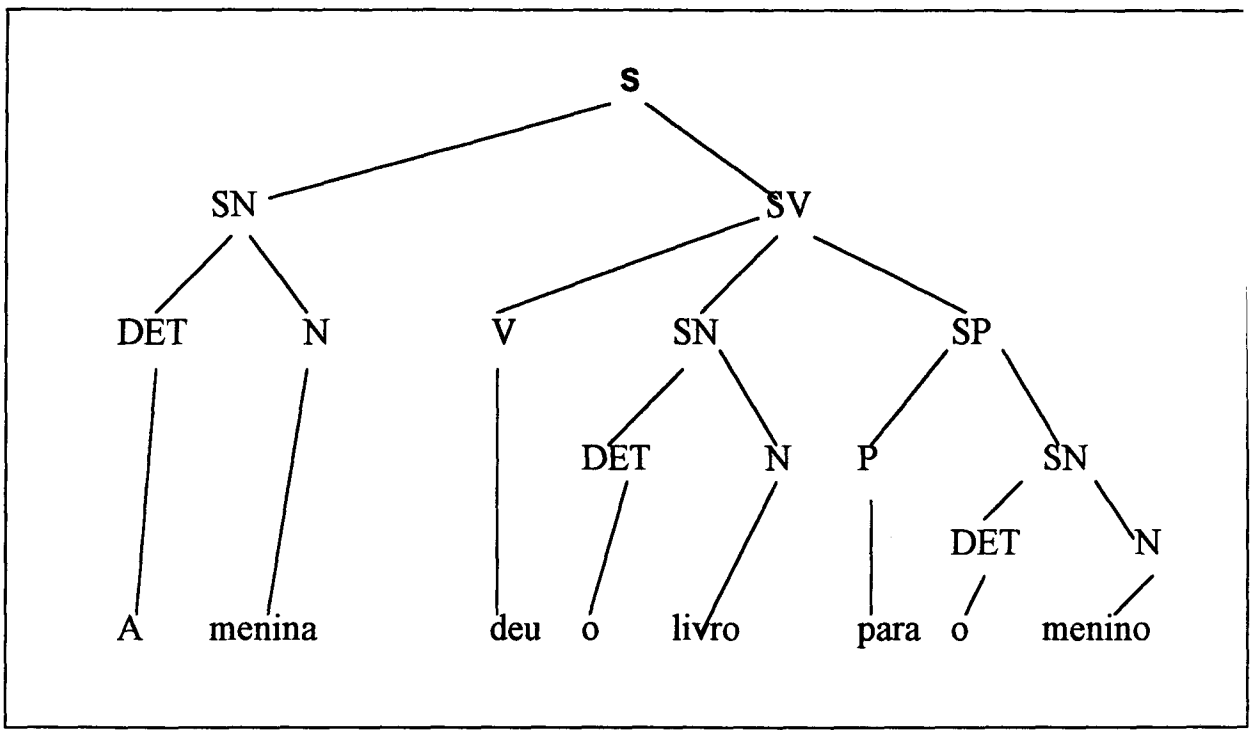
Noam Chomsky, distribucionalista de formação. Chomsky definia a língua como o conjunto infinito das sentenças enunciadas ou enunciáveis pelos falantes, e postulava a existência de dois níveis de descrição da estrutura sintática: o nível superficial (***s-structure***) e o nível profundo (***d-structure***). Para Chomsky, a regularidade lingüística que se observava na superfície das sentenças era antes a manifestação de uma regularidade que operava em um nível mais profundo, cujo acesso somente poderia se dar pela exploração da intuição lingüística dos falantes. Nascia, portanto, o conceito de **competência lingüística**, que seria exatamente aquilo que habilitaria os falantes a lidar com a produtividade da língua, ou seja, a entender e a produzir enunciados por eles nunca antes ouvidos ou produzidos. De acordo com Chomsky, essa competência lingüística poderia ser expressa por um conjunto finito de regras (relacionado à *d-structure*) que, operando sobre o vocabulário da língua, regularia a produção dos enunciados lingüísticos (a *s-structure*). Nesse conjunto finito de regras estariam compreendidos um componente de base — formado por **regras gerativas**[17] — e um conjunto de **transformações**, que produziriam alterações na configuração da estrutura profunda[18]. A coexistência de regras gerativas e transformações levaria o formalismo proposto por Chomsky a ser freqüentemente referenciado como **gerativo-transformacional**[19].

A revolução chomskyana se torna particularmente pertinente para o processamento automático das línguas naturais porque traz, como subproduto, um formalismo gramatical, de natureza lógica, teoricamente capaz de descrever todo o conjunto de sentenças de uma determinada língua. Trata-se da ***phrase-structure grammar – PSG*** (ou estrutura de constituintes imediatos, ou estrutura de marcadores frasais, ou estrutura sintagmática, segundo as traduções correntes para o português do Brasil), que é formalmente definida como a quádrupla <T,N,P,S>, em que T representa o vocabulário terminal (as palavras da língua); N, o vocabulário não-terminal (as categorias funcionais e as categorias lexicais da língua); P, o conjunto de regras de produção (ou regras de reescrita); e S, o símbolo inicial, membro de N. Do ponto de vista prático, o formalismo convida à representação de sentenças como estruturas arbóreas invertidas, que têm o

[17] A regra que permite que o símbolo inicial (S) se reescreva como sintagma nominal (SN) e sintagma verbal (SV) é um exemplo de regra gerativa: <S> ::= <SN> <SV>

[18] A regra que, atuando sobre a estrutura ativa, permitiria a produção de sentenças passivas é um exemplo de transformação.

símbolo inicial como raiz, as categorias lexicais e funcionais como ramos, e o vocabulário da língua como folhas, cuja distribuição superficial seria o resultado dos mecanismos gerativos e transformacionais que se aplicariam para a estrutura profunda.



< S > ::= <SN> <SV>
<SV> ::= <V> < SN> <SP>
<SN> ::= <DET> <N>
<SP> ::= <P> <SN>
<V> ::= deu
<P> ::= para
<DET> ::= o | a
<N> ::= menino | menina | livro[20]

Figura 5.1. Representação e gramática para a sentença *A menina deu o livro para o menino.*

A configuração das regras de produção das PSGs governaria, segundo Chomsky, o seu poder descritivo. O autor estabelece uma hierarquia de gramáticas (que pressupõe uma hierarquia de línguas) a partir do número e da natureza de símbolos que ocupam as

[19] Para uma abordagem menos esquemática da teoria gerativa, o leitor deve consultar os originais de Chomsky, particularmente (Chomsky, 1957; 1965; 1986).

posições das regras de produção (Chomsky, 1959). Seriam quatro as variações possíveis das PSGs:

- Gramáticas do tipo 3, ou gramáticas regulares (**regular grammars**), ou gramáticas de estados finitos (**finite-state grammars**), cujas regras se adequariam a uma de três formas: <A> :: = <B>t | t<B> | t, em que <A> e <B> são símbolos não-terminais, e t é um símbolo terminal;
- Gramáticas do tipo 2, ou gramáticas livres de contexto (**context-free grammars**), cujas regras obedeceriam à sintaxe <A> :: = x, em que x pode ou não ser um símbolo terminal;
- Gramáticas do tipo 1, ou gramáticas sensíveis ao contexto (**context-sensitive grammars**), cujas regras seriam da forma x ::= y, em que o comprimento de y é maior ou igual ao comprimento de x;
- Gramáticas do tipo 0, ou gramáticas irrestritas (**unrestricted grammars**), cujas regras não seguiriam qualquer padrão.

Segundo Chomsky, as gramáticas de tipo 0 serviriam à descrição de qualquer (tipo de) língua, mas seu excessivo poder descritivo seria de pouca utilidade na compreensão dos fenômenos lingüísticos, porque estariam contempladas, na gramática, mesmo as sentenças que não pertencem à língua que se pretende descrever. O ideal seria a utilização de formalismos gramaticais menos poderosos, capazes de produzir apenas as sentenças efetivamente aceitáveis para uma determinada língua. No caso da maior parte das línguas naturais, incluído o português, acredita-se que as gramáticas livres de contexto sejam as mais adequadas.

A *phrase-structure grammar* originalmente proposta por Chomsky sofrerá, com o tempo, uma série de modificações ou aumentos, voltados para a representação de funções não previstas na proposta original. O conjunto de transformações sofreria, durante a década de 1960, uma série de alterações, sendo por fim abandonado por uma única regra de movimento (*move-α*) no início da década de 1970. A interveniência de categorias semânticas também produziria alterações no modelo original, com a introdução da teoria temática e de estratégias de representação do conteúdo semântico

[20] A sintaxe das regras acompanha aqui a notação BNF (Backus-Naur Form), em que os símbolos não-terminais são representados entre "<" e ">", e o símbolo de reescrita é "::=". A barra vertical "|" marca as diferentes possibilidades de reescrita.

lexical. Com o tempo, a concepção do símbolo inicial seria problematizada pela teoria x-barra, levando à substituição de S por projeções das categorias funcionais da própria sentença. A própria integridade da representação estrutural foi colocada à prova, e a fragmentação da estrutura sintática levou à construção de formalismos de unificação. Algumas dessas alterações foram produzidas pela própria teoria gerativa, que conheceu, desde o seu surgimento, diferentes orientações teóricas. Assim, a Gramática Padrão (*Standard Theory*, ou simplesmente ST), a Gramática Padrão Estendida (*Extended Standard Theory*, EST) e a Teoria da Regência e Ligação (*Government and Binding Theory*, GB) constituem, apenas no âmbito do gerativismo, três diferentes momentos de revisão do mesmo formalismo gramatical (*Phrase-Structure Grammar*). Concorrentemente, proliferarão, nas décadas de 1980 e 1990, outros modelos teóricos que introduzirão novas estratégias de formalização gramatical. Entre os principais representantes dessas novas estratégias, dissidentes do gerativismo, estão a gramática léxico-funcional (*Lexico-Functional Grammar*, LFG) e a *Generalized Phrase-Structure Grammar* (GPSG), que tem hoje na *Head-Driven Phrase-Structure Grammar* (HPSG) seu principal representante. Infelizmente não pode pertencer ao escopo deste trabalho, flagrantemente introdutório, o aprofundamento de cada uma dessas vertentes gramaticais, ficando o leitor orientado para a consulta da bibliografia complementar[21].

5.4 As gramáticas

Os formalismos abordados no item anterior conformam um princípio de descrição das sentenças da língua, e não a sua própria descrição. Não se deve confundir a sintaxe das regras com as regras propriamente ditas. Dizer que o português pode ser descrito por uma gramática livre de contexto não significa dizer que exista apenas uma gramática livre de contexto capaz de descrever o português. Os formalismos gramaticais, como arquitetura do conjunto de regras, serão preenchidos por regras específicas, derivadas das mais diferentes correntes teóricas.

No entanto, percebe-se que, diferentemente do que ocorre na gramática tradicional, a teoria lingüística não tem se preocupado com a elaboração de modelos gramaticais completos, robustos, capazes de descrever todas as sentenças que compõem o português. Talvez por influência do gerativismo, que pretende descrever não línguas

[21] Sobre a LFG, o leitor deve consultar (Bresnan, 1982). Duas abordagens interessantes sobre GPSG estão em (Sampson, 1983) e (Gazdar et al., 1985).

naturais específicas, mas os princípios universais que governam a faculdade da linguagem, talvez pela insuficiência dos modelos propostos, não existem gramáticas formais exaustivas para a língua portuguesa. Poderão ser encontradas descrições de fenômenos genéricos, relativos à língua, mas em nenhum momento se estabelece um modelo total, ainda que imperfeito, capaz de processar sentenças em tempo real. Esta constitui seguramente uma das principais limitações no desenvolvimento de ferramentas computacionais para a língua portuguesa, porque a sintaxe, como se verá na próxima seção, desempenha um papel-chave no processamento automático das línguas naturais.

5.5 A importância da sintaxe para o PLN

O papel da componente sintática nas línguas naturais é controverso, variando entre a posição central a ela atribuída pela Teoria Padrão e a posição marginal a ela consignada pelos semântico-gerativistas. Para os primeiros, a semântica é uma interpretação da sintaxe; para os últimos, a sintaxe é uma projeção da semântica. A verdade — admite-se hoje — talvez esteja no meio-termo. O conhecimento sintático revela-se (a) dispensável em estruturas que apresentam alto grau de previsibilidade semântica (como em "*Ele vai cinema amanhã mulher*", em que a presença das preposições é quase desnecessária), e (b) imprescindível em estruturas não tão corriqueiras (como "*O menino mordeu o cachorro*", em que a ausência de sintaxe conduziria a uma interpretação exatamente inversa àquela que se pretenderia transmitir). De qualquer forma, reconhece-se, de maneira geral, que ainda que o nível semântico e o nível sintático possam envolver algum grau de redundância, a compensação dos ruídos inevitáveis no processo de comunicação somente se torna possível se estiverem disponíveis pistas das duas naturezas.

Na lingüística computacional, contudo, a questão assume um viés diferente. Ainda não estão disponíveis estratégias satisfatórias de representação do conhecimento semântico para o computador. Toda a teoria lingüístico-semântica se estrutura a partir dos conceitos de referência (ou *denotatum*) e sentido (ou *designatum*), ainda irrepresentáveis para a máquina, que não pode perceber o mundo (identificando, assim, os referentes das formas lingüísticas), nem formar, a partir dele, uma imagem psíquica. De resto, a interveniência de fatores alheios ao co-texto lingüístico na produção do sentido, como as variantes contextuais (os atos de fala, as implicaturas conversacionais, a dêixis, etc.) fartamente assinaladas pela pragmática lingüística, torna praticamente

irreplicável o comportamento semântico observado para as línguas naturais. Impõe-se, portanto, para a lingüística computacional, como conseqüência da intratabilidade dos fenômenos semânticos, a centralidade da componente sintática, pelo menos até que se desenvolvam outras estratégias de representação do significado lingüístico.

5.6. O parsing[22]

O processamento automático da sentença com o objetivo do reconhecimento de sua estrutura sintática recebe tradicionalmente o nome de ***parsing***. Por extensão, a ferramenta que executa esse conjunto de procedimentos [que permite assinalar funções sintáticas a cada um dos itens lexicais que compõem a sentença] é referenciada como ***parser***. A história tem revelado que os *parsers* podem variar de acordo com (a) a relação que estabelecem com o usuário; (b) os recursos disponíveis; e (c) as estratégias de análise. Nesta seção, passaremos em revista essas três perspectivas de abordagem.

Da relação com o usuário

Aqui se repete o que normalmente acontece em toda a prática lingüístico-computacional: o grau de automatismo das ferramentas é variável. Os *parsers* podem ser completamente automáticos, realizando solitariamente toda a análise sintática, todo o processo de desambigüização lexical e sintática, todo o processo de reconhecimento das sentenças da língua. E os *parsers* podem recorrer eventualmente ao usuário, diante de construções inesperadas, diante de ambigüidades insolúveis, na ausência de estratégias de decisão. Nos dois casos, a concepção do *parser* dependerá de sua finalidade e da disponibilidade e da competência metalingüística do usuário. Em ferramentas de tradução automática ajudada (*machine-aided translation systems*, MAT), é esperável que haja algum diálogo entre a máquina e o tradutor, particularmente para o provimento das referências contextuais de que a máquina não dispõe. Por outro lado, ferramentas de correção gramatical (*grammar checkers*) geralmente dispensam a ajuda do usuário, cujo domínio dos princípios gramaticais está sendo posto em xeque. Em um e outro caso, o grau de dependência do usuário está diretamente relacionado à quantidade e à qualidade dos recursos disponíveis.

[22] Para uma análise mais detalhada, ainda introdutória, do *parsing* o leitor deve consultar o capítulo segundo de (Grishman, 1986) e os capítulos sexto e sétimo de (Smith, 1991).

Um outro princípio de classificação dos *parsers* remete ao número de análises geradas. Neste caso, os *parsers* podem ser **probabilísticos** ou **determinísticos**. Serão determinísticos quando consignarem apenas uma estrutura sintática à sentença analisada, escolhendo, com ou sem a ajuda do usuário, uma entre várias estruturas concorrentes; serão probabilísticos quando apresentarem, para a mesma sentença, todas as possibilidades de análise sintática, geralmente hierarquizadas segundo alguma probabilidade de ocorrência. Nos dois casos, a natureza da ferramenta será determinada pela aplicação.

Da relação com os recursos disponíveis

A análise sintática automática das sentenças de uma língua natural traz algumas exigências incontornáveis, como a disponibilidade de um léxico e de uma gramática que antecipem as formas verificáveis nas sentenças que se pretende analisar. Da estruturação desse léxico e dessa gramática poderão emergir novas necessidades, como um conjunto de estratégias de regularização e de desambigüização lexical e sintática.

O **dicionário** a ser acessado pelo *parser* pode assumir formatos variados: pode ser apenas uma lista de itens (morfemas, palavras, locuções, expressões, frases inteiras) ou uma estrutura composta de formas (sub)categorizadas. A complexidade do *parser* será inversamente proporcional à quantidade e à qualidade das informações presentes no dicionário. A associação das formas a categorias lexicais (principalmente a informação relativa às partes do discurso) é normalmente tomada como requisito mínimo para o processamento sintático automático, já que as gramáticas formais geralmente tomam as categorias lexicais como a última instância dos símbolos não-terminais. No entanto, a desambigüização dos casos de homonímia pode requerer informações mais refinadas, geralmente relacionadas à explicitação das valências sintáticas ou do conteúdo semântico dos verbetes.

O princípio de classificação lexical deve evitar a ambigüidade, sob o risco de proliferação das possibilidades de análise sintática. No entanto, nem sempre é possível precisar, no próprio dicionário, as relações semânticas e sintáticas de verbetes homógrafos. O processamento sintático requererá, nesses casos, a aplicação de estratégias de desambigüização categorial. A **desambigüização lexical** visa a evitar a explosão combinatória derivada da ambigüidade das informações representadas no

léxico. Quanto mais ambíguas as categorias lexicais, ou mais numerosos os casos de homografia, tanto mais necessários os princípios de desambigüização. Esses princípios geralmente são regulados por critérios estatístico-distribucionais. Considera-se a probabilidade de ocorrência de determinada forma a partir de fatores variados: o co-texto lingüístico (à esquerda e à direita) e suas restrições selecionais; ou o contexto de uso, com suas variantes de forma e de conteúdo (ou domínio).

Além do dicionário, o funcionamento do parser está diretamente relacionado à disponibilidade de uma **gramática**, ou de um conjunto de regras (ou de princípios) que permita à máquina testar a gramaticalidade (a boa formação gramatical) das sentenças da língua. Fosse a língua um conjunto finito de sentenças, gramáticas não seriam necessárias. Bastaria dicionarizar as estruturas lingüísticas: listaríamos todas as sentenças da língua, as analisaríamos sintaticamente e armazenaríamos os resultados para permitir comparações futuras. O reconhecimento da estrutura de uma sentença não passaria, portanto, de uma função de acesso a um banco de dados previamente compilado. Embora uma análise acurada do uso da língua revele que são extremamente comuns as construções fixas, formulaicas, as frases prontas (como os provérbios ou os clichês), é forçoso reconhecer que as sentenças das línguas naturais constituem antes um conjunto aberto, infinito, marcado pela heterogeneidade da forma.

Heterogeneidade não significa, porém, irregularidade, e a análise, mesmo superficial, de um conjunto representativo de sentenças da língua permitirá encontrar padrões de comportamento sintático razoavelmente recorrentes. Em português, por exemplo, o artigo precede o substantivo, e o adjetivo concorda com o substantivo que ele modifica. São regras que se depreendem do uso da língua e que já foram recuperadas por qualquer gramática normativa do português. Uma estratégia para o desenvolvimento de um *parser* seria, pois, investi-lo desse conhecimento já explicitado sobre os processos de formação das sentenças da língua. Em outras palavras: deveríamos ensinar ao *parser* tudo aquilo que as gramáticas sabem. Evidentemente, como já foi assinalado na terceira seção deste capítulo, seria preciso antes matematizar as regras gramaticais para que elas pudessem ser manipuladas pelo computador. Isso normalmente é feito através de variações de *phrase-structure grammars*, o modelo formal proposto por Noam Chomsky. A precedência do artigo sobre o substantivo poderia ser representada para a

máquina como uma regra do tipo: <SN> ::= <DET> <N>, que poderia ser processada a partir de um algoritmo simples, como o que se segue:

```
x = 0;
leia posição(x);
enquanto posição(x) for diferente do marcador de fim de sentença:
        se posição(x) = artigo,
                então leia posição(x+1);
                        se posição(x+1) = substantivo,
                          então  posição(x) = determinante;
                                 posição(x+1) = núcleo do sintagma nominal;
                                 distância(x,x+1) = sintagma nominal;
                caso contrário,
                        não é uma sentença bem-formada da língua portuguesa;
        x = x + 1;
        leia posição(x).
```

Evidentemente, trata-se de um algoritmo incompleto e de uma versão bastante simplificada do que é o processamento sintático, que deve considerar um sem-número de outras variantes, não previstas na regra esquemática proposta, exclusivamente dedicada à identificação do artigo que precede o substantivo. No entanto, é forçoso reconhecer que já estamos falando, nesse nível, a língua do computador, que poderia, a partir da implementação do algoritmo, identificar alguns dos sintagmas nominais que compõem as sentenças da língua.

O problema é que nem todas as regras de boa formação sintática são conhecidas. Sabe-se que artigo precede o substantivo, que o sujeito concorda com o verbo, mas em nenhuma parte se encontra uma análise exaustiva das formas que o sujeito pode assumir na língua portuguesa. Que o núcleo do sujeito deve ser uma expressão de natureza substantiva (um substantivo, um pronome substantivo, um adjetivo substantivado, uma oração subordinada substantiva) é certo; mas as nuanças que cada uma dessas formas pode adquirir na realização efetiva do sujeito não mereceram ainda consideração sistemática ou resposta definitiva da teoria lingüística. A análise das sentenças abaixo permite observar com alguma clareza as dificuldades que a matéria encerra:

(1) A alegria e o contentamento era enorme.

(2) Aconteceu um acidente terrível na estrada.

(3) Alguém sempre sai ganhando.

(4) Vendem-se casas.

(5) Ele desapareceu.

(6) O pessoal foram no cinema.

(7) ∅ dizem que a inflação vai voltar.

(8) Flores não tem acento.

(9) Falta dois dias pra acabar o ano.

(10) Fumar provoca câncer.

(11) ∅ comprei um carro novo.

(12) Mais de um deputado votou contra a proposta.

(13) Mateus, Marcos, João e Lucas foram apóstolos de Jesus Cristo.

(14) O príncipe dos sociológos virou presidente.

(15) O quiabo desapareceu dos supermercados.

(16) O menino que vimos ontem passeando na rua quando estávamos a caminho do teatro desapareceu.

(17) Os Lusíadas é um livro de Luís de Camões.

(18) Walter Benjamin se matou.

(19) Paulo saiu de casa e ∅ desapareceu.

(20) Sair de casa, em São Paulo, à tarde, durante o mês de março, quando o céu está cinzento, é pedir para ficar preso na chuva.

(21) Não faça ∅ isso, Maria!

Todos os termos grifados nas sentenças acima exercem aquilo a que se convencionou chamar função de sujeito. É interessante observar que a sua determinação não é tão simples quanto faz parecer a gramática normativa. O sujeito possui forma extremamente heterogênea, dimensão variável, nem sempre concorda com o verbo (caso de 1, 6, 8, 9 e 17), nem sempre vem anteposto ao verbo (caso de 2, 4 e 9), pode ser indeterminado (caso de 7) ou elíptico (caso de 11, 19 e 21). Encontrar uma regularidade subjacente a essa aparente diversidade de forma, tamanho, posição e uso não é tarefa simples, e freqüentemente envolve um grau de conhecimento sobre a língua que ainda não foi atingido.

A estratégia mais utilizada nestes casos é submeter a sentença a um pré-processamento, antes da execução do *parsing*. Trata-se do processo de **regularização sintática**, através do qual as informações omitidas (os sujeitos elípticos, por exemplo) são restauradas, as anáforas são indicializadas, as formas passivas são substituídas pelas formas ativas, a sentença é reorganizada a partir da ordem direta (sujeito verbo objeto), e as clivagens e topicalizações são suprimidas. Enfim, operam-se transformações, no sentido chomskyano do termo, para que a heterogeneidade seja reduzida. No entanto, colocam-se algumas dificuldades: o impacto da regularização pode por vezes afetar o sentido da sentença, trazendo implicações semânticas sérias para o processamento da língua (é o caso, por exemplo, da substituição das formas passivas que contêm quantificadores); em muitos casos, as estratégias de regularização (como a indicialização das anáforas) dependem do processamento sintático, e não podem, portanto, precedê-lo; por fim, a regularização sintática é incapaz de restaurar relações extra-sentenciais (co-textuais ou contextuais) fundamentais para a reorganização da sentença.

Percebe-se, portanto, que a estratégia de dotar o *parser* das regras que se encontram nas gramáticas da língua portuguesa possui alcance limitado. Em primeiro lugar, porque não existe uma gramática definitiva e tampouco a certeza de que algum dia ela possa vir a ser elaborada (principalmente se considerarmos que a língua é um fenômeno social, que varia incessantemente no tempo, no espaço, nas camadas da sociedade). Em segundo lugar, porque muitos dos critérios de boa formação sintática (os critérios de gramaticalidade) talvez não sejam matematizáveis. A subjetividade interfere no julgamento, e a boa formação das sentenças pode depender de fatores ligados à cooperatividade do falante, como a atenção e a motivação, por exemplo. Por fim, a pretensa sistematicidade das sentenças da língua portuguesa cai por terra na análise das frases produzidas no registro oral, marcadas por falsos inícios, hesitações, repetições, retomadas, anacolutos, topicalizações e movimentos de natureza pouco previsível e de regularidade bastante discutível.

Construir uma gramática da língua portuguesa, capaz de descrever todas as sentenças possíveis, não é, portanto, tarefa trivial. O procedimento padrão tem sido a composição de gramáticas específicas a subdomínios da língua ou a categorias funcionais inferiores à sentença, o que fragiliza o caráter robusto do *parsing*. Em vez de analisar quaisquer

sentenças, os *parsers* normalmente analisam partes da sentença (como o sintagma nominal) ou sentenças pertencentes a domínios específicos (como o registro da escrita na norma culta da língua), cujo grau de previsibilidade é consideravelmente maior do que o da totalidade das formas possíveis em português.

Das estratégias de análise

De posse de um léxico e de uma gramática, o *parser* pode começar a análise propriamente dita, que consiste na recuperação das funções sintáticas desempenhadas pelos itens lexicais da sentença, e pela consignação, à sentença, de uma estrutura sintagmática hierarquizada. Neste percurso, a análise poderá se dar de várias formas: da esquerda para a direita, da direita para a esquerda, de cima para baixo, de baixo para cima, ou de forma combinada. A escolha dos movimentos depende em grande medida do tipo de gramática adotado.

Em sentido horizontal, a estratégia de análise mais comum é a que obedece à linearidade da língua, que vai da esquerda para a direita. Esta constitui a hipótese mais realista do ponto de vista psicológico. Na fala, como na escrita, os humanos não esperamos o fim da sentença para começarmos a processá-la. O processamento é feito em tempo-real, o que restringe a possibilidade de que os procedimentos de análise possam percorrer a direção contrária (da direita para a esquerda) do movimento da língua.

Em sentido vertical, predomina o processamento *top-down*, de cima para baixo, partindo do símbolo inicial para a construção da sentença. Há também aqui certo realismo psicológico, atestado pelas antecipações que os humanos normalmente fazemos no processamento das sentenças. Uma outra virtude do processamento *top-down* é a possibilidade de recursividade, que reduz o conjunto de regras da gramática. No entanto, essa mesma recursividade envolve problemas de controle (principalmente no caso da recursão à esquerda) que podem afetar seriamente o desempenho da ferramenta. O processamento das sentenças pode enveredar por labirintos sintáticos dos quais o *parser* somente consegue sair após um número excessivamente dispendioso de *backtrackings*. Outro problema característico da abordagem *top-down* é o seu caráter tudo-ou-nada: ou traçamos toda a estrutura sintática da sentença ou não identificamos

nenhuma das estruturas sintagmáticas parciais que a compõem, por mais que possam ser previstas pela gramática utilizada.

A alternativa, particularmente neste último caso, é a análise *bottom-up*, de baixo para cima, que parte das categorias lexicais para chegar às categorias funcionais. O problema aqui são as regras de generalização, que permitem a identificação das fronteiras sintagmáticas. Sem a visão do conjunto, a identificação das fronteiras se torna um problema de solução nada trivial, que geralmente envolverá a análise da sentença por núcleos, exigindo pois um formalismo de unificação: constroem-se, primeiramente, estruturas parciais e, a partir da aplicação de regras de combinação dessas estruturas, chega-se à estrutura de toda a sentença. Esse tipo de análise envolve geralmente a disponibilidade de duas gramáticas: uma gramática que opera sobre categorias lexicais e categorias funcionais nucleares e outra que opera sobre projeções de categorias funcionais.

No meio termo, entre as estratégias de análise *top-down* e *bottom-up*, estão os parsers híbridos (como os *chart parsers*), que apostam em uma combinação dos dois movimentos. Essa combinação pode se dar de forma paralela, quando se disparam dois *subparsers*, operando em direções opostas, cujas convergências serão fixadas; ou de forma seqüencial, quando um movimento de análise (*top-down*, por exemplo) é suspenso até que a ferramenta tenha dados suficientes do movimento contrário para tomar decisões. Nos dois casos, o formalismo gramatical deverá contemplar as especificidades do modelo de análise, produzindo novamente uma gramática para cada estratégia de processamento.

5.7 Comentários Finais

É preciso salientar que as limitações aqui expostas, derivadas da complexidade da matéria e da incipiência dos estudos a ela relativos, não significam, muito pelo contrário, a impossibilidade ou a inutilidade da análise sintática automática das sentenças das línguas naturais. Haverá um conjunto de sentenças, bastante significativo, que pode e deve ser tratado a partir da construção de gramáticas formais como as descritas acima. Ainda que não se possa chegar a um modelo total da língua, aproximações poderão ser atingidas que se revelam mais úteis do que inúteis. O revisor gramatical desenvolvido no NILC, que será apresentado no próximo capítulo, é prova

de que o tratamento da língua portuguesa, ainda que fragmentário e ainda que simplificado, pode ajudar o usuário humano em sua interação pela língua.

6. Ferramentas e Aplicações

Ao longo das aulas anteriores, uma série de tópicos foi discutida em que abordagens lingüísticas, computacionais ou lingüístico-computacionais foram introduzidas, além de ser feita menção a possíveis aplicações. Neste capítulo, pretendemos retomar o tópico de ferramentas e/ou recursos lingüísticos para o processamento de línguas naturais (PLN), com ênfase nos trabalhos desenvolvidos no NILC. Ao usar tais projetos como exemplo de aplicações e para ilustrar a necessidade de recursos lingüísticos, aproveito a oportunidade para descrever nossa pesquisa recente. Além disso, discutirei aspectos de gestão de projetos científico-tecnológicos em uma área multidisciplinar, que podem servir de elemento de reflexão sobre as dificuldades e desafios desse tipo de projeto. Faço a ressalva que, neste último tópico, a discussão será baseada em minha visão da área, que pode ser bastante diferente até de meus próprios colegas do NILC. Como será também comentado, é comum haver discrepâncias entre abordagens a serem seguidas, o que torna o processo decisório um grande desafio.

6.1. O Projeto ReGra

Este projeto nasceu do interesse da Itautec/Philco em contar com ferramentas de correção para o seu editor de textos, Redator. O primeiro contato ocorreu em 1993, e o ponto de partida das discussões foi a possibilidade de se desenvolver um revisor ortográfico e gramatical, semelhante àqueles disponíveis então para o inglês. A equipe do NILC, criado como um núcleo informal, havia sido indicada à Itautec/Philco porque tínhamos docentes no Departamento de Computação e Estatística com formação em PLN (Nunes, 1991; Aluísio, 1989) e alguma experiência em ferramentas de auxílio à escrita, no projeto AMADEUS (Aluísio & Oliveira Jr., 1995) para confecção de texto científico em inglês. Os primeiros meses serviram para a identificação dos tópicos de pesquisa a serem abordados inicialmente, e a formação de uma equipe que contasse com lingüistas e cientistas da computação. Uma análise desse trabalho inicial, com nossa experiência atual, mostra que o caminho trilhado para desenvolver pesquisas tecnológicas, a partir de experiências puramente acadêmicas que até então era o que possuíamos, e de caráter multidisciplinar, requer um grande investimento na formação de uma equipe. O investimento não é só financeiro, para poder atingir pluralidade

através de profissionais de áreas diferentes, mas também de trabalho de aprendizado para o estabelecimento de uma linguagem comum.

6.1.1 Concepção e Arquitetura do ReGra

Chamamos de ReGra o sistema de correção gramatical, não incluindo as rotinas para detecção de erros ortográficos, embora a base lexical que suporta o corretor ortográfico tenha sido compilada para o projeto de correção gramatical. O ReGra é constituído por três módulos principais: i) o módulo estatístico, ii) o mecânico e iii) o módulo gramatical. As rotinas para compactação e acesso aos dados do léxico foram desenvolvidas pela equipe do Prof. Tomasz Kowaltowski, do Instituto de Informática da Unicamp (Kowaltowski & Lucchesi, 1993).

O módulo de tratamento estatístico realiza uma série de cálculos, fornecendo parâmetros físicos de um texto sob análise, como o número total de parágrafos, sentenças, de palavras, de caracteres, etc. O componente mais importante desse módulo, entretanto, é o que fornece o "índice de legibilidade" (Martins et al., 1996), uma indicação do grau de dificuldade da leitura do texto. O conceito de índice de legibilidade surgiu a partir do trabalho de Flesch (Flesch, 1948) para a língua inglesa e busca uma correlação entre tamanhos médios de palavras e sentenças e a facilidade de leitura. Não inclui aspectos de compreensão do texto, que requereriam tratamento de mecanismos complexos de natureza lingüística, cognitiva e pragmática. O índice Flesch, assim como outros similares, tem sido empregado para uma grande variedade de línguas, mas o trabalho do NILC foi o primeiro para a língua portuguesa. Através de um estudo comparativo de textos originais em inglês e traduzidos para o português, verificou-se que a equação que fornece o índice Flesch precisaria ter seus parâmetros adaptados para o português, pois as palavras desta língua são em média mais longas, em termos do número de sílabas, do que em inglês.

A adaptação do índice Flesch para o português resultou na identificação de quatro faixas de dificuldade de leitura, conforme indicado na Tabela I.

Tabela I – Faixas para o índice de Flesch modificado

Índice Flesch modificado	Grau de Dificuldade
75 a 100	Muito fácil
50 a 75	Fácil
25 a 50	Difícil
0 a 25	Muito difícil

Textos classificados como **muito fáceis** seriam adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental; textos **fáceis** seriam adequados a alunos com escolaridade até a oitava série do ensino fundamental; textos **difíceis** seriam adequados para alunos cursando o ensino médio e/ou universitário, e textos **muitos difíceis** em geral seriam adequados apenas em áreas acadêmicas específicas. Por se tratar de um dado estatístico, o índice de legibilidade só é calculado para trechos com mais de 100 palavras. Testes realizados com textos tradicionalmente dirigidos a públicos dessas quatro faixas mostraram resultados bastante satisfatórios. Por exemplo, jornais de grande circulação como a Folha de São Paulo e o Estado de São Paulo têm em seus cadernos principais índices de legibilidade que correspondem a textos adequados a leitores com escolaridade equivalente ao final do ensino fundamental. Textos de cadernos infantis, por outro lado, apresentam índices de Flesch modificados na faixa de muito fácil, ou seja, podem em princípio ser acompanhado por crianças que ainda não completaram os quatro primeiros anos do ensino fundamental.

O segundo módulo do ReGra, o mecânico, detecta erros facilmente identificáveis que não são percebidos por um corretor ortográfico. Exemplos desse tipo de erro são: i) palavras e símbolos de pontuação repetidos; ii) presença de símbolos de pontuação isolados; iii) uso não balanceado de símbolos delimitadores, como parêntesis e aspas; iv) capitalização inadequada, como o início da sentença com letra minúscula; v) ausência de pontuação no final da sentença.

O módulo gramatical, obviamente o mais importante, é tratado numa seção à parte, que se segue.

apontam desvios nas relações entre núcleos e adjuntos, entre núcleos e modificadores, entre regentes e regidos. A realização de análise sintática automática obviamente requer que todas os itens lexicais estejam categorizados apropriadamente. Para tanto, realizou-se em paralelo a construção do léxico, que envolveu a compilação exaustiva das palavras da língua portuguesa e a hierarquização das categorias dos itens lexicais morfologicamente ambíguos.

Uma vez que alguns erros em contextos lingüísticos específicos ocorrem independentemente de desvios sintáticos, na versão atual do ReGra convivem as duas abordagens mencionadas acima. Ou seja, além de realizar análise sintática automática, muitas das regras heurísticas da primeira versão foram mantidas, como as de correção de erros de crase.

O desempenho do Revisor, quanto a tempo de execução, pode ser considerado como ótimo, uma vez que as mensagens de erro são apresentadas ao usuário, praticamente, instantaneamente. As limitações do Revisor, entretanto, estão localizadas nos falsos erros que ainda comete e nos erros não detectados. A maior parte destes problemas advém da impossibilidade de serem previstas todas as estruturas sintáticas desviantes que podem ser empregadas por usuários médios. Embora o sistema conte, hoje, com mais de 600 produções, ainda aparecem estruturas para as quais nenhum casamento (matching) é obtido. Além disso, as dificuldades advindas de pluricategorização de alguns itens lexicais, principalmente nos casos de homonímia, precisarão ser tratadas em casos especiais, o que certamente demandará grandes esforços de pesquisa. Em algumas situações, a inserção de conhecimento semântico no léxico é indispensável, sendo essa uma meta da nossa equipe para o futuro próximo. Vários tópicos lingüísticos relevantes para o desenvolvimento do ReGra foram relatados em (Martins et al., 1998).

A Figura 6.1 mostra um diagrama de blocos ilustrativo do ReGra. São representados o módulo de configuração, em que as regras de correção e estilo podem ser habilitadas ou não, o módulo mecânico já mencionado anteriormente, e o módulo gramatical. Mostra-se também que o ReGra trata parágrafos individualmente cujos componentes – em termos de palavras e símbolos – são identificados no analisador léxico. Quanto ao módulo gramatical, ressalta-se a presença de regras de correção pontuais e baseadas na análise sintática.

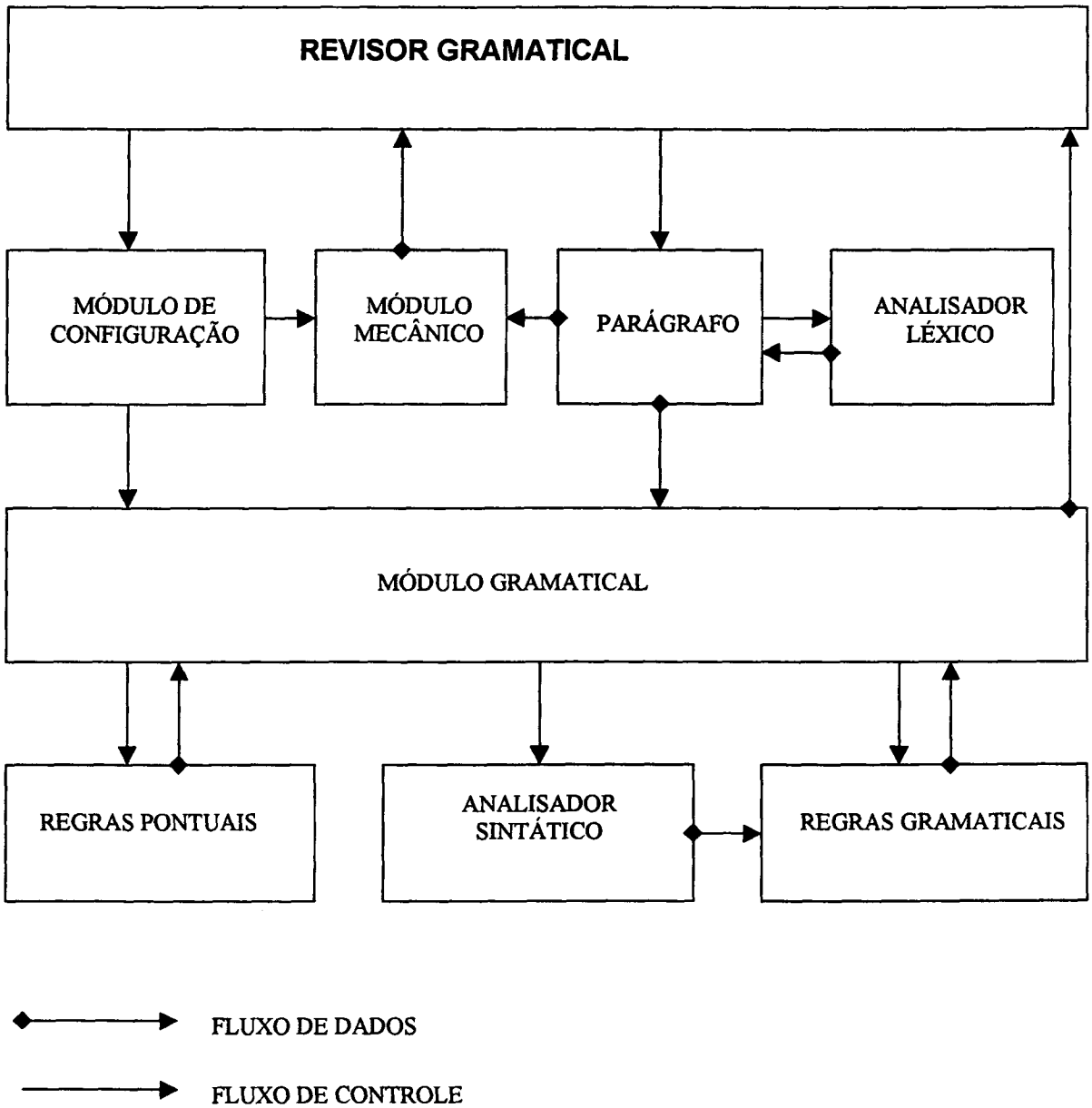


Figura 6.1. Arquitetura do ReGra

6.1.3. Exemplos de erros detectados pelo ReGra

A elaboração das regras de correção teve como base um estudo de gramáticas contemporâneas do português do Brasil. As regras foram testadas em textos autênticos, sem correção, com o objetivo de verificar sua operação e a eventual ocorrência de falsos erros. A seguir ilustramos algumas classes de erros detectados pelo Revisor.

Crase

Só recorrerei à essa alternativa em último caso.

Chegou à conclusões variadas.

Costumava ler o evangelho durante às refeições.

A boa safra começa à partir de julho deste ano.

Ausência de crase

Eu vou **as** duas horas ao encontro marcado. .

O carro parou devido **a** falta de combustível.

O problema referente **a** economia não pertence à lei do inquilinato.

Colocação Pronominal

Me deu um presente.

Nunca **vi-a** tão gorda.

Eu **darei-te** todo o auxílio que puder.

Se tivesse dinheiro, daria-**lhe** um bom presente.

Uso do pronome

Eu fiquei fora de **si**.

Ele viu **ele**.

Nós se preocupamos.

As visitas bateram à porta. Mande **elas** entrar. .

Reduziram **os aumentos salariais.** Reduziram-**os**. .

Concordância Verbal de Particípio

Foram aprovado todas as alunas.

Foi apontado a inconveniência de se promover uma festa.

Foi detectado uma pane no sistema de ar-condicionado.

Concordância Nominal e Verbal

Ele foram para a escola.

A mulher e a menina ficaria amigas.

Todos informações eram imprecisas.

A maioria dos corredores chegaram ao fim da prova.

Coloque ponto final **aos** final **da** sentenças.

As palestra de Laudelino pelo Brasil afora podem ter fim.

Os fogos de artifício da noite de São João são **lindo**.

A guerra entre os deputados pelas possíveis alianças **começaram**. .

Deu três horas **no** relógio da matriz. .

Há uma semana, acabou **as férias**.

Começou **as aulas** no novo colégio de ensino médio e fundamental da cidade.

Tudo é flores. .

Deve**mos** nos manter**mos** em pé.

Inadequações no uso dos verbos Fazer e Haver

Ele chegou **a** dois anos.

Fazem dois meses que não tomo cerveja.

A muito tempo moro nesta casa. .

Vou visitá-la daqui **há** dois dias. .

Naquele ano **houveram** poucos acontecimentos que valem a pena recordar.

A Partícula "Se"

Vende-se casas.

Precisam-se de funcionários.

Expressões Fixas

haja visto, cujo(s) o(s), bancas de jornais, toalhas de mesas

Prefixos

neo-clássico

auto afirmação

semianalfabeto.

Regência

Eu assisto **o** jogo, **o** filme e **a** novela. .

Aonde você está? .

Onde você vai? .

Prefiro escrever **do que** falar. .

Uso do particípio regular/irregular

Os criminosos **foram pegados** pela polícia.

A polícia **tem pego** criminosos.

Pontuação

Os meninos prodígios da cidade do interior, ficaram na capital.

O interesse no trabalho informal no Brasil, cresceu a partir dos anos 90.

Vícios de linguagem

Ele reincidiu **de novo** no erro.

Ele subiu **para cima** do palco.

Os alpinistas desceram **para baixo** da montanha.

Inadequação lexical

A rua é **melhor iluminada**. .

A moça comprou **duzentas** gramas de ameixas.

Emprego de mau/mal

O **mal** filho não saiu de casa.

Mau cheguei e já tenho que sair.

6.1.4. Recursos Lingüísticos

(a) Léxico

O léxico compilado no projeto de colaboração com a Itautec/Philco serviu para os revisores ortográfico e gramatical. Uma descrição detalhada da compilação desse léxico pode ser encontrada em (Nunes et al., 1996). Para o revisor ortográfico, o léxico deve ser o mais abrangente possível, contendo inclusive nomes próprios, siglas, abreviaturas, etc. Já para o módulo gramatical, as palavras do léxico precisam ser categorizadas quanto a sua classe gramatical, o que dificulta a manipulação de grandes massas de dados requeridas pela abrangência do revisor ortográfico. A compilação de um conjunto de palavras para a formação de um léxico é conceitualmente simples, apesar do enorme volume de trabalho envolvido. De fato, a compilação do presente léxico tomou praticamente um ano de trabalho de três lingüistas e dois informatas, dedicando-se respectivamente 30 e 20 horas semanais, atuando no ICMC-USP de São Carlos. Foram necessários, também, o trabalho de digitadores e a cooperação eventual de colaboradores, principalmente com vistas à adequação do léxico para o sistema de correção gramatical. Além disso, o que em princípio parecia um trabalho mecânico, ainda que exaustivo, acabou mostrando facetas interessantes com perspectivas de uma nova gama de pesquisas em lexicografia.

Partindo-se de um conjunto de aproximadamente 120 mil palavras normalmente encontradas em dicionários impressos, o maior trabalho consistiu em expandir o conjunto com: a) as conjugações dos verbos, b) as flexões de gênero, c) as flexões de número, d) as derivações de grau. Essas tarefas foram todas feitas automaticamente, a partir de algoritmos formulados pelos lingüistas. Para a maioria dessas tarefas foi necessária uma revisão "manual" cuidadosa para a detecção de malformação de palavras. Ressalte-se que a decisão de se construir um léxico cujas entradas são palavras (no máximo, palavras compostas hifenizadas) deveu-se a duas razões básicas: 1) o suporte ao revisor ortográfico não permitiria ou, pelo menos, dificultaria o uso de um

léxico baseado em regras de aglutinação de morfemas, uma vez que, ao permitir construções morfologicamente válidas, estaríamos permitindo o uso de palavras que configurariam erros de fato (p.ex. *imexível*); 2) o conjunto inicial de verbetes foi extraído de um dicionário eletrônico, o que certamente economizou tempo e esforços. Essa decisão implica, entre outras coisas, que expressões que se queira considerar como *tokens* únicos, como as locuções, devem ser manipuladas num contexto extraléxico. Neste caso, a saída é indicar, de forma *ad hoc* no léxico, os prováveis componentes de expressões que, por sua vez, devem estar disponíveis na forma de listas (caso das locuções) ou mesmo na forma de programas (caso dos nomes próprios, em que se consideram ocorrências consecutivas de nomes próprios como um único nome próprio).

Testes do revisor ortográfico empregando o léxico (parcial) construído a partir do conjunto de verbetes inicial mostraram um desempenho insuficiente. Por isso, adicionalmente às formas previstas, um grande trabalho de verificação de formas faltantes foi feito utilizando-se um corpus que conta, hoje, com aproximadamente 37 milhões de palavras. Através desse trabalho com o corpus, o léxico foi expandido e atualmente conta com cerca de 1.500.000 lexemas gerados a partir de aproximadamente 100.000 lemas.

Para um bom desempenho do revisor gramatical, problemas de outra natureza aparecem na construção do léxico. O conjunto de atributos de cada palavra no léxico varia em relação à categoria principal da palavra, e engloba as seguintes informações: categoria gramatical (substantivo, verbo, adjetivo, pronome, artigo, numeral, preposição, advérbio, conjunção, nome próprio, sigla, abreviatura,...), e, dependendo de cada caso de categoria, gênero, número, grau, predicação, regência (nominal/verbal), tipo (de adjetivo, de conjunção, etc.), tempo, pessoa, colocação pronominal (ênclise, mesóclise). Toda entrada tem uma forma canônica associada, que permite relacionar todas as entradas (lexemas) que possuem uma forma comum, ou seja, um mesmo lema. Por exemplo, menino, meninos, menina, meninas têm em comum o lema menino, e dessa forma é possível recuperar as diferentes flexões a partir de cada um deles. Neste aspecto, deparamo-nos com situações particulares, como em ouros que tanto pode ser um lexema derivado de ouro, como o lema ouros, referindo-se ao naipe do baralho. Assim, há duas entradas ouros, ambas de mesma categoria sintática (substantivo), porém cada uma com sua forma canônica distinta (ouro e ouros). Outro problema

relacionado às canônicas é ilustrado pelas variantes parasito/parasita. Se considerarmos um único lema, parasito, o revisor deixaria de aceitar a forma "o parasita". Assim, optamos por associar as próprias formas adjetivas como suas canônicas.

Dois problemas para a classificação dos itens lexicais foram a pluricategorização dos lexemas e o tratamento de homônimos. No estágio atual, o léxico contém entradas desprovidas de qualquer significação, não permitindo distinções gramaticais da língua que requerem conhecimento semântico, como na homonímia. Para ilustrar, *cedo* pode ser tanto advérbio como uma forma conjugada do verbo *ceder*. Obviamente, não há problema em considerar as duas classificações. Mas, dependendo da consulta a ser feita pelo revisor gramatical, haverá de ser fornecida a classificação "mais provável" da palavra. Como os revisores são de propósito geral, para a hierarquização das possíveis categorias decidiu-se adotar o critério de freqüência de uso. Essa tarefa, no entanto, não é simples devido à indisponibilidade de dicionários de freqüência para a língua portuguesa. Uma vez que o corpus, embora extenso, não tem equilíbrio de representatividade quanto a tipos de textos, apelamos para nossa intuição de falante e estudiosos da língua, e, posteriormente, confrontamos os resultados com os dados de freqüência do corpus.

A experiência adquirida na construção do léxico mostrou a necessidade de se dispor de um corpus representativo da língua em uso. Com a possibilidade de empregar ferramentas de software para lidar com estas grandes massas de dados, abrem-se novas perspectivas de pesquisas em lexicografia, e mesmo construção de novos tipos de dicionários. Para isso, muito contribuirá a geração automática de novas palavras (por exemplo: advérbios terminados em "mente", adjetivos com sufixo "ável", palavras justapostas como em "interdiscurso"), que obedeçam às regras de formação de palavras para o português. É óbvio que a verificação, por um especialista, das palavras geradas é essencial, como já ocorreu na compilação do nosso léxico. Além disso, levando-se em conta a freqüência de uso, é possível criar dicionários adequados para um dado grupo de usuários, incluindo palavras de uso freqüente geradas automaticamente e que em geral não constam dos dicionários impressos, e evitando-se palavras que jamais são empregadas por aquele grupo-alvo. Outra possibilidade é a criação de dicionários técnicos. Uma busca no corpus pode não só identificar os termos técnicos mais freqüentes, mas também apontar estrangeirismos que se incorporam à língua por falta de

similar em português e mesmo termos técnicos "aportuguesados", muitas vezes sem o cuidado de obedecer às regras de formação de palavras em português.

A inserção de itens não dicionarizados, incluindo nomes próprios, siglas, etc. deve ser feita com bastante critério, e este é um tópico que tem gerado muita discussão no NILC. As discussões em geral giram em torno do estabelecimento de critérios confiáveis e consistentes para justificar a inclusão de tais itens. Minha perspectiva é a do usuário: um grande número de intervenções desnecessárias, causadas pela ausência de nomes, termos técnicos, etc., prejudica sobremaneira a utilização do ReGra. Defendo, portanto, que o léxico seja o mais abrangente possível, embora não se possa perder de vista a adequação dos itens inseridos.

(b) Corpus

Por várias vezes foi mencionada a importância de serem realizados testes com o ReGra, para verificar seu desempenho do ponto de vista da precisão na correção, inclusive sobre a incidência de falsos erros e omissões. Testes realísticos só podem ser obtidos se textos autênticos forem empregados. Aqui, deve-se frisar a necessidade de variedade de textos. Por exemplo, é essencial que textos não corrigidos sejam usados em testes para verificar se o revisor de fato detecta erros comuns. Por outro lado, o revisor não pode ser concebido de maneira a intervir desnecessariamente com grande freqüência, o que requer testes em textos sem erros gramaticais para simular a utilização da ferramenta por um usuário que escreve corretamente. Em vista dessas necessidades, o corpus contém textos de livros científicos, literários e jornalísticos, com predominância para este último tipo, por questão de disponibilidade. Não houve preocupação em garantir representatividade do corpus quanto às diferentes tipologias de texto do português do Brasil, mas sim reunir um banco de textos para testes. O corpus conta hoje com cerca de 37 milhões de palavras.

Ainda com relação a testes, é importante poder comparar diferentes versões de um revisor gramatical, e mesmo testar seu desempenho de acordo com diferentes critérios, como a freqüência de falsos erros e omissões. Para tais testes mais específicos, foi criado um corpus artificial no sentido de que foram selecionadas sentenças de vários tipos: i) contendo erros detectados pelo ReGra, ii) contendo erros não detectados pelo ReGra (omissões), iii) contendo casos em que o ReGra comete um falso erro.

Para a pesquisa em um corpus tão extenso, deve-se contar com ferramentas de busca específicas. No NILC, temos empregado um conjunto de ferramentas criadas por um grupo de pesquisa alemão de Stuttgart, além de algumas outras simples para verificar freqüência de itens lexicais em arquivos de texto.

6.2 O Projeto UNL

O grande volume de informações disponíveis na Internet continua restrito a um número muito pequeno de usuários, principalmente nos países cuja língua não é o inglês. A importância dessa restrição talvez ainda não seja muito sentida entre nós, pois apenas uma pequena porcentagem da população tem acesso à Internet. E é a parcela da população que, em geral, tem maior probabilidade de conhecer, pelo menos instrumentalmente, a língua inglesa. Mas a tendência é a de que a Internet se popularize muito rapidamente, atingindo várias camadas da população, como hoje acontece com televisores, videocassetes e até TV a cabo. A partir do momento em que grande parte da população tiver acesso à Internet, como garantir que ela possa desfrutar de todo o potencial da rede? É impossível ensinar inglês para toda essa massa de gente. Ou seja, um mecanismo de tradução terá que ser utilizado para que as pessoas possam receber - e transmitir - informações em sua própria língua.

Devido ao volume de informações, a tradução humana pode ser descartada. Não só pelo custo altíssimo que teria, mas também porque a velocidade com que se realiza a tradução impediria que a informação fluísse com a rapidez necessária. A alternativa seria então uma tradução automática, realizada por um programa de computador. Já existem no mercado alguns softwares que realizam essa tarefa, para várias línguas. Para o português em particular, são mais comuns os que traduzem textos do inglês para o português, mas a possibilidade de tradução no sentido inverso dificilmente é oferecida.

Sensível à enorme barreira da língua para a comunicação entre os povos, a Universidade das Nações Unidas (UNU), sediada em Tóquio, resolveu patrocinar um projeto de longa duração - 10 anos - para o desenvolvimento de ferramentas de software para vencer essa barreira. Nesse projeto, ao invés de tradução de uma língua para outra, faz-se a codificação do conteúdo de um texto em uma dada língua natural em uma língua artificial, chamada *Universal Networking Language* (UNL), criada especificamente para textos escritos no ambiente da Internet. O texto já codificado em UNL pode então ser decodificado para a

língua destino. O projeto UNL é de âmbito mundial, sendo coordenado pelo Instituto de Estudos Avançados, da Universidade das Nações Unidas. Nos três primeiros anos, a partir de 1997, estão sendo desenvolvidos codificadores e decodificadores para cerca de 15 línguas, incluindo chinês, russo, alemão, francês, italiano, japonês, inglês, hindi, espanhol e português. Nos 7 anos restantes previstos para o projeto UNL, espera-se atingir praticamente todas as línguas oficiais das Nações Unidas. As ferramentas para o português estão sendo desenvolvidas no Brasil, sob a coordenação do Prof. Tadao Takahashi, com a participação do NILC.

O que distingue o projeto UNL é a tentativa de conjugar esforços de grupos de pesquisa de vários países. Para que haja essa colaboração, estão sendo realizados congressos específicos com participantes dos diversos países, ocasiões nas quais são debatidos avanços e dificuldades no desenvolvimento das ferramentas, inclusive com sugestões de extensões para a língua artificial, UNL. Deve-se ressaltar, também, que há características bastante diferenciadas neste projeto em comparação com similares em tradução automática envolvendo várias línguas. Além de empregar a abordagem de interlíngua, o desenvolvimento é distribuído – inclusive geograficamente – com especialistas trabalhando em suas próprias línguas.

6.2.1. Usando a interlíngua UNL

A representação da sentença em UNL ilustra um conjunto de relações cujos argumentos compõem o vocabulário da UNL. A expressão que vem antes dos parênteses (tim, nam, ppl, etc.) sinaliza a relação semântica entre as expressões dentro dos parênteses. As expressões demarcadas por "@" (entry, pred, past, etc.) indicam os atributos gramaticais dos argumentos envolvidos na relação semântica. A UNL inclui atualmente cerca de 35 relações semânticas. O léxico da UNL, que em princípio é uma versão sem ambigüidades de um dicionário inglês, contém uma parte genérica, ou seja, as entradas de um dicionário inglês (begin, city, build, etc.), e uma parte específica, de acordo com a categoria da palavra, ou seja, de sua posição na ontologia subjacente à UNL. Por exemplo, em São Paulo(icl>city), refere-se à palavra "São Paulo", pertencente à categoria de cidade, isto é, a cidade de São Paulo.

A interlíngua UNL propicia uma representação única para o conteúdo semântico de uma sentença de uma língua natural. É, portanto, uma metalíngua para descrever aspectos

especiais do significado da sentença, como as relações semânticas que podem ser representadas através de relações formais (morfológicas ou sintáticas) entre palavras de uma sentença. É capaz de representar o significado de orações, mesmo sem contar com um modelo de gramática. Embora bastante limitada no escopo de suas relações, a UNL traz a vantagem de permitir uma representação do conteúdo semântico que não exige conhecimento da estrutura profunda da sentença. A UNL se restringe a tratar o significado literal das sentenças, sem levar em conta aspectos como estilo, intenção, retórica, etc., característicos do significado conotativo.

Os ingredientes da UNL são: palavras universais (*Universal Words*, UWs), rótulos de relação (*Relation Labels*, RLs) e rótulos atributivos (*Attribute Labels* (ALs). A função de uma UW é expressar um significado específico. Cada UW é representada por uma palavra inglesa que contém o significado genérico da língua inglesa, acompanhada de restrições de maneira que essa mesma palavra inglesa gere várias UWs. Conseqüentemente, ambigüidades devido à homografia são eliminadas. Por exemplo, a palavra "book" tem várias UWs a ela associadas, para indicar os seus vários significados, como em book(icl>publication), ou seja, livro, ou book(obj>room), ou seja, reservar um quarto. Portanto, a UW "book" contém todos os significados possíveis, ao passo que as outras com restrições servem para desambigüização. O trabalho de associação de palavras do português às UWs fornecidas pela Universidade das Nações Unidas está descrito em (Dias-da-Silva et al,; 1998).

Os RLs expressam relações binárias entre significados, ou seja, UWs. Sua representação genérica é um par do tipo *relation_label*(UW_1, UW_2), onde UW_1 e UW_2 são inter-relacionadas através da relação semântica denotada pelo RL. Há várias classes de RLs que podem ser associadas a dois componentes da sentença, ou 2 UWs. Um exemplo é a relação (agente-objeto), em que o agente (agt) é um objeto animado que causa uma ação volitiva. Outros RLs incluem método (met), tempo (tim), beneficiário (ben), possessão (pos). Os RLs também são empregados para relacionar UWs na restrição de significados, como inclusão (icl) para representar hiperonímia, como em *icl(dog, animal)*, ou sinônimo (*equ*) para representar significados equivalentes entre UWs.

Os rótulos atributivos (ALs) são empregados para restringir o significado genérico de uma UW. Informações como tempo, aspecto e intenção também são representadas como

ALs. Um AL é representado por uma UW seguida de atributos identificados pelo símbolo "@". Por exemplo, uma UW com *n* atributos tem a forma: UW.@attrib1.@attrib2.....@attribn. Se nenhum AL for associado a uma UW, esta terá o significado mais geral de sua classe. Assim como para os RLs, há diferentes tipos de ALs. Alguns tipos mais comuns são: i) os que restringem UWs, ii) os que expressam tempo verbal; iii) os que expressam aspecto, intenção, etc. Os ALs do tipo i) podem ser, por exemplo, *@pl* e @generic, que indicam, respectivamente, uma UW no plural e de caráter genérico. Este é o caso da sentença *Peter eats apples*, representada em UNL por:

agt(eat.@present,Peter),
obj(eat.@present,apple.@generic.@pl).

Outros ALs deste tipo são @def, *@indef* e *@not*. Já os ALs que expressam tempo verbal foram definidos de acordo com a gramática inglesa, incluindo *@past*, *@present*, e *@future*. Quanto aos ALs que expressam aspecto, são exemplos *@begin-soon*, "evento que irá começar ", como em *The airplane is about to land*, representado por:

agt(land@begin-soon,plane.@def).

Outros exemplos de ALs que expressam aspecto são *@begin-just*, *@end-soon*, *@end-just*, *@progress* para eventos que estão em curso, *@repeat* para a repetição de um dado evento que geralmente envolve agente/objeto. Os ALs para expressar intenção incluem: *@emphasis*, *@topic*, *@intention*, *@recommendation*, etc. Alguns são aplicados em componentes intra-sentenciais, enquanto outros se aplicam a uma sentença completa.

6.2.2. Decodificação de UNL para o português

No processo de codificação, a UNL permite um mapeamento das sentenças de uma dada língua natural num conjunto de relações entre significados, sem representação explícita da estrutura sintática. No processo de decodificação, por outro lado, regras gramaticais da língua destino são aplicadas à sentença em UNL para gerar uma sentença em língua natural. Um sistema de regras foi desenvolvido no NILC para o português do Brasil, que permite a decodificação de sentenças relativamente complexas (Martins et al., 1998). O sistema funciona acoplado a uma ferramenta, denominada DECO, fornecida pela Universidade das Nações Unidas.

A Figura 6.2 abaixo mostra o funcionamento do decodificador. Uma sentença codificada em UNL é introduzida no decodificador, que emprega regras de geração de sentenças em português a partir da representação semântica em UNL. Obviamente, as UWs originais da sentença UNL precisam ser traduzidas para o português, e para tanto se emprega o dicionário UW-português.

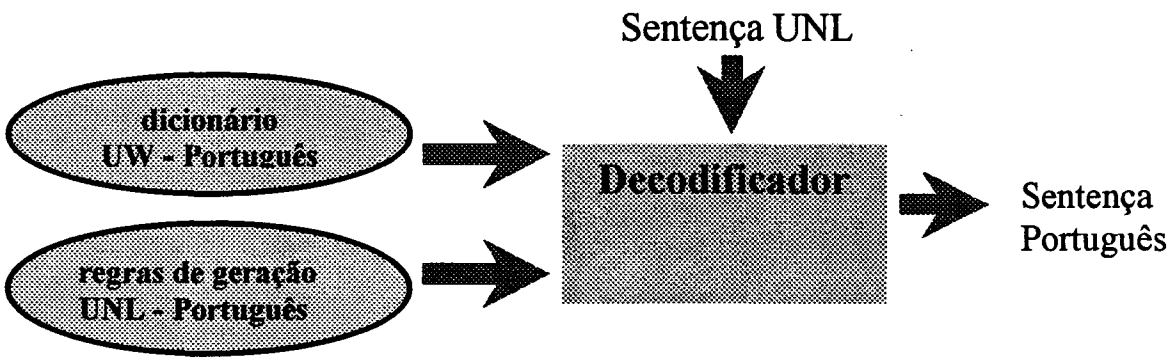


Figura 6.2. Arquitetura do DeCodificador UNL-Português

Segue o exemplo de uma sentença original inglês, codificada manualmente em UNL, e posteriormente decodificada automaticamente para o português.

Sentença original:

It shall function in accordance with the annexed Statute, which is based upon the Statute of the Permanent Court of International Justice and forms an integral part of the present Charter.

Sentença decodificada:

A corte funcionará de acordo com o estatuto anexo que se baseia no estatuto da corte permanente de justiça internacional e constitui uma parte integrante da carta presente.

A título de ilustração, é apresentada a representação UNL para essa sentença.

```
obj(function(icl>event).@entry.@pred.@obligation,
      court(icl>judiciary place):01.@def)
man(function(icl>event)@entry.@pred.@obligation,
      in accordance with(icl>manner))
obj(in accordance with(icl>manner), statute(fld>law):01.@def)
aoj(annexed, Statute(fld>law):01.@def)
obj(base(icl>event).@pred, statute(fld>law):01.@def)
bas(base(icl>event).@pred, statute(fld>law):02.@def)
```

mod(statute(fld>law):02.@def, court(icl>judiciary place):02.@def)
aoj(permanent(icl>state), court(icl>judiciary place):02.@def)
mod(court(icl>judiciary place):02.@def, justice(equ>judiciary))
aoj(international(icl>state), justice(equ>judiciary))
and(form(equ>constitute).@pred, base(agt>organization,icl>set, ppl>place).@pred)
obj(form(equ>constitute).@pred, statute(fld>law):01.@def)
gol(form(equ>constitute).@pred, part(icl>quantity).@indef)
aoj(integral(icl>state), part(icl>quantity).@indef)
mod(part(icl>quantity).@indef, charter(icl>document).@def)
aoj(present(icl>state), charter(icl>document).@def)

6.3. Comentários Finais

Gostaria de comentar alguns aspectos de gestão de projetos de natureza multidisciplinar, com objetivos tanto acadêmicos quanto tecnológicos. Em primeiro lugar, é importante ressaltar que o PLN para o português é bem menos desenvolvido do que para outras línguas comercialmente importantes, principalmente o inglês. Para o português, é necessário, ainda, criar uma série de recursos lingüísticos como corpora anotados, dicionários eletrônicos processáveis automaticamente, analisadores sintáticos e semânticos robustos, sem o qual é difícil realizar pesquisas que possam gerar bons resultados científicos e também aplicações práticas. A grande dificuldade para a obtenção desses recursos lingüísticos advém do fato que muito desse tipo de trabalho não gera resultados acadêmicos, como dissertações e teses. É importante, assim, poder contar com recursos de fontes diferentes das tradicionais: em outras palavras, tem-se que buscar recursos financeiros na iniciativa privada, o que por sua vez requer o desenvolvimento de "produtos", passíveis de comercialização.

A necessidade de dirigir pelo menos parte da pesquisa para produtos é bastante restritiva. Entretanto, há vantagens também na medida em que os pesquisadores são forçados a trabalhar com um domínio mais próximo possível da realidade, o que acaba garantindo a realização de pesquisas básicas em que o estudo da língua em uso é essencial. A maior dificuldade está na formação de uma equipe interdisciplinar, principalmente no Brasil onde não existe um curso que forme profissionais, em lingüística ou computação, cujo perfil seja orientado especificamente à Lingüística Computacional.

Agradecimentos: Os autores agradecem a toda a equipe do NILC, sem a qual essa iniciativa não seria possível.

Referências

ALLEN, J. (1995). *Natural Language Understanding*. The Benjamin/Cummings Pub. Co. 2ª ed.

ALUÍSIO, S.M. (1989) *Tratamento de Ambiguidade de Escopo de Quantificadores em Processamento de Linguagem Natural,* Dissertação de Mestrado, ICMC-USP.

ALUÍSIO, S.M.; OLIVEIRA Jr., O.N. (1995) A case-based approach for developing writing tools aimed at non-native English users, *Lecture Notes in Artificial Intelligence*, 1010, 121.

APPELT, D. (1985). *Planning Natural Language Utterances*. Studies in Natural Language Processing. Cambridge University Press.

BECHARA, E. (1972). *Moderna gramática portuguesa*. São Paulo: Companhia Editora Nacional. p. 73.

BRESNAN, J. (1982). *The mental representation of grammatical relations*. Cambridge, MA: The MIT Press.

CAMARA JR., J.M. (1989) *Princípios de lingüística geral*. Rio de Janeiro: Padrão.

CHOMSKY, N. (1957). *Syntactic Structures*. The Hague/Paris: Mouton.

CHOMSKY, N. (1959). On certain formal properties of grammars. In *Information and Control 2*, 137-167.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass: The MIT Press.

CHOMSKY, N. (1986). *Knowledge of language – its nature, origin and use*. Westport/London: Praeger.

CLOCKSIN, W. and MELLISH, C. (1981). *Programming in Prolog*. Springer-Verlag, New York.

COLMERAUER, A. (1977). *An Interesting Subset of Natural Language*. Groupe Intelligence Artificielle, Faculté des Sciences de Luminy, Marseille, France.

CUNHA, C. & CINTRA, L. (1985) *Nova Gramática do português contemporâneo*. Rio de Janeiro: Nova Fronteira.

DALE, R.; MELLISH, C.S.; ZOCK, M. (eds.) (1990). *Current Research in Natural Language Generation.* Academic Press.

DALE, R. (1992). *Generating Referring Expressions.* ACL-MIT Press Series in Natural Language Processing, Cambridge, Ma.

DIAS-DA-SILVA, B.C. (1997) Bridging the gap between linguistic theory and natural language processing. In: Bernard Caron (ed.) *Proceedings of the 16th International Congress of Linguists*, **16**, 1997, Paris. Anais..., Oxford: ELSEVIER SCIENCE-PERGAMON, 1998, Paper 0425, ISBN 0 08 043 438X

DIAS-DA-SILVA, B.C.; SOSSOLOTE, C.; ZAVAGLIA, C.; MONTILHA, G.; RINO, L.H.M.; NUNES, M.G.V.; OLIVEIRA JR., O.N.; ALUÍSIO, S.M. (1998) The Design of a Brazilian Portuguese Machine Tractable Dictionary for an Interlingua Sentence Generator, III Encontro para o Processamento Computacional do Português Escrito e Falado, Porto Alegre, RS.

FLESCH, R.(1948) A new readability yardstick, J. Appl. Psychology, 32, 221-233.

GAZDAR, G. et alii. (1985). *Generalized Phrase-Structure Grammar*. Oxford: Basil Blackwell.

GRICE, H.Paul (1975). Logic and Conversation. In P. Cole and J.L. Morgan (eds.), *Syntax and Semantics. Volume 3: Speech Acts*, pp. 41-58. Academic Press, New York.

GRISHMAN, R. (1986). *Computational Linguistics – an introduction.* Cambridge: Cambridge University Press.

GROSZ; Barbara J.; SPARCK JONES, Karen and WEBBER, Bonnie Lynn (eds.) (1986), *Readings in Natural Language Processing*, Morgan Kaufmann Publishers, Inc. California.

HOVY, E. (1988). *Generating Natural Language under Pragmatic Constraints.* Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

KOWALSKI, R. (1974). *Logic for Problem Solving.* Memo No. 75. Dept. of Computational Logic, University of Edinburgh, Edinburgh, UK.

KOWALTOWSKI, T.; LUCCHESI, C.L. (1993) Applications of finite automata representing large vocabularies. *Software-Practice and Experience,* 23(1), 15-30.

MARTINS, R.T.; RINO, L.H.M.; NUNES, M.G.V.; OLIVEIRA JR., O.N. (1998) Can the syntactic realization be detached from the syntactic analysis during generation of natural language sentences?, III Encontro para o Processamento Computacional do Português Escrito e Falado, Porto Alegre, RS.

MARTINS, R.T.; HASEGAWA, R.; NUNES, M.G.V.; MONTILHA, G.; OLIVEIRA Jr., O.N.(1998) Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. *Natural Language Engineering*, 4 (4) 287-307.

MARTINS, T.B.F.; GHIRALDELO, C.M.; NUNES, M.G.V.; OLIVEIRA Jr., O.N. (1996). Readability formulas applied to textbooks in Brazilian Portuguese, *Notas do ICMSC-USP*, Série Computação, 28.

MANN, W.C.; THOMPSON, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190, June 1987.

MATTHIESSEN, C.M.I; BATEMAN, J.A. (1991). *Text Generation and Systemic-Functional Linguistics*, Pinter Publishers, London.

McDONALD, D.D.; BOLC, L. (eds.) (1988). *Natural Language Generation Systems.* Springer-Verlag, New York, NY.

McKEWON, K.(1985). *Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Studies in Natural Language Processing. Cambridge University Press.

McKEWON, K.; SWARTOUT, W.R. (1987). Language Generation and explanation. *The Annual Review of Computer Science,* (2):401-449.

MINSKY, M. (1975). A Framework for Representing Knowledge. In P. Winston (ed.), *The Psychology of Computer Vision*, MacGraw-Hill, New York.

NUNES, M.G.V. (1991) *A Geração de Respostas Cooperativas em Sistemas Baseados em Lógica.* Tese de Doutoramento, PUC-Rio.

NUNES, M.G.V. et alii. (1996) A Construção de um Léxico para o Português do Brasil: Lições Aprendidas e Perspectivas. Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado. CEFET-PR, Curitiba.

PARIS, C.; SWARTOUT, W. and MANN, W.C. (eds.) (1991). *Natural Language Generation in Artificial Intelligence and Computational Linguistics.* Kluwer Academic Publishers, Boston.

PERINI, M. A. (1996) *Gramática descritiva do português.* São Paulo: Ática.

_____ (1976) *Gramática gerativa – introdução ao estudo da sintaxe portuguesa.* Belo Horizonte: Vigília.

QUILLIAN, M.R. (1968). Semantic Memory. In *SIP*, pp. 216-270.

RICH, E.; KNIGHT, K. (1993). *Inteligência Artificial.* 2ª ed. Makron Books, São Paulo.

RUMELHART, D.E. and NORMAN, D.A. (1975). The Active Structural Network. In D.A. Norman and D.E. Rumelhart (eds.), *Explorations in Cognition.* W.H. Freeman, San Francisco.

SAMPSON, G. R. (1983). Context-free parsing and the adequacy of context-free grammar. In KING, M. (ed.). *Parsing Natural Language*. London: Academic Press.

SAUSSURE, F. de. (1988). *Curso de Lingüística Geral*. São Paulo: Cultrix.

SCHANK, R.C. and ABELSON, R.P. (1977). *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Assoc., Hilssdale, NJ.

SHIEBER, S.M. (1984). The Design of a Computer Language for Linguistic Information". In *Proc. of the 10th International Conference on Computational Linguistics - COLING'84;* Stanford University, CA. pp.362-366.

SHIEBER, S.M. (1986). An Introduction to Unification-based Approaches to Grammar. *CSLI Lecture Notes*, Vol. 4. University of Chicago Press, 1986.

SHIEBER, S.M.; PEREIRA, F.; KARTTUNEN,L.; KAY, M. (1986). *A Compilation of Papers on Unification-Based Grammar Formalisms*. CSLI Report No. CSLI-86-48.

SIMMONS, R.F. (1973). Semantic Networks: Their Computation and Use for Understanding English Sentences. In R. Schank and K. Colby (eds.), *Computer Models of Thought and Language*, pp. 63-113. W.H. Freeman, San Francisco.

SMADJA, F.A.; McKEOWN, K.(1991). Using collocations for language generation. *Computational Intelligence*, 7(4):229-239, December.

SMITH, G. W. (1991). *Computers and Human Language*. Oxford: Oxford University Press.

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

WINSTON, P.H. (1993). *Artificial Intelligence*. 3rd ed. Addison-Wesley.

WOODS, W.A. (1986). Transition Network Grammars for Natural Language Analysis. In Barbara J. Grosz; Karen Sparck Jones and Bonnie Lynn Webber (eds.), *Readings in Natural Language Processing*, Morgan Kaufmann Publishers, Inc. California.