Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

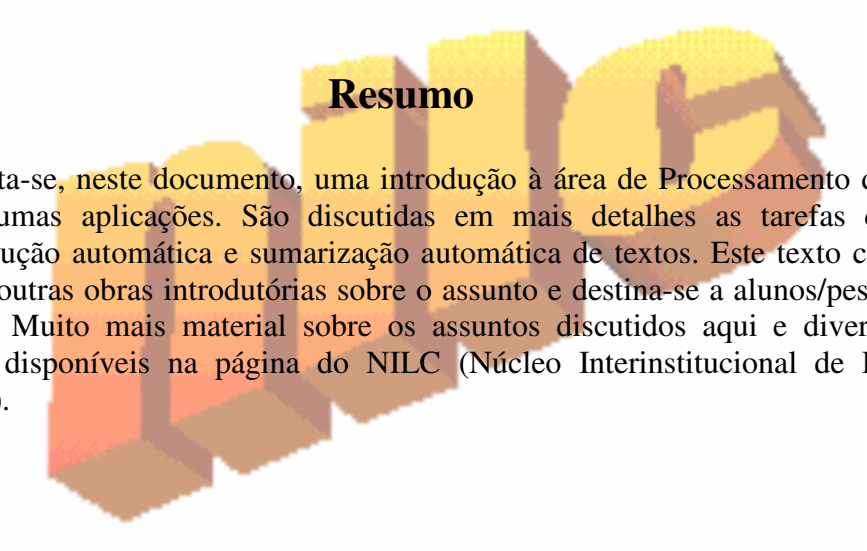
Introdução ao Processamento das Línguas Naturais e Algumas Aplicações

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

NILC-TR-07-10

Agosto, 2007

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil



Resumo

Apresenta-se, neste documento, uma introdução à área de Processamento de Línguas Naturais e algumas aplicações. São discutidas em mais detalhes as tarefas de revisão gramatical, tradução automática e sumarização automática de textos. Este texto consiste em uma junção de outras obras introdutórias sobre o assunto e destina-se a alunos/pesquisadores novos na área. Muito mais material sobre os assuntos discutidos aqui e diversos outros assuntos estão disponíveis na página do NILC (Núcleo Interinstitucional de Lingüística Computacional).

Esse documento foi produzido pela junção das seguintes obras (disponíveis na página do NILC):

- Nunes, M.G.V.; Dias da Silva, B.C.; Rino, L.H.M.; Oliveira Jr., O.N.; Martins, R.T.; Montilha, G. (1999). *Introdução ao Processamento das Línguas Naturais*. Notas Didáticas do ICMC, N. 38. São Carlos/SP, Junho, 91p.
- Specia, L. e Rino, L.H.M. (2002). *Introdução aos Métodos e Paradigmas de Tradução Automática*. Serie de Relatórios Tecnicos do NILC, NILC-TR-02-04. São Carlos/SP, Março, 23p.
- Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação*, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, pp. 203-245. Campinas/SP. Agosto.

ÍNDICE

PARTE I – Introdução ao Processamento das Línguas Naturais

1. Introdução ao PLN

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, uma introdução aos diferentes tipos de conhecimento lingüístico para o tratamento de línguas naturais, a apresentação de arquiteturas de sistemas de interpretação e geração de línguas naturais, uma introdução ao processo automático de análise sintática, tão importante na maioria das aplicações de PLN, e a apresentação do projeto ReGra, no qual vários conceitos apresentados anteriormente serão ilustrados. 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 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.

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

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 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

[2] (Winograd, 1972)

> 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:

> "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.

3 Grifo nosso.

4 Grifo nosso.

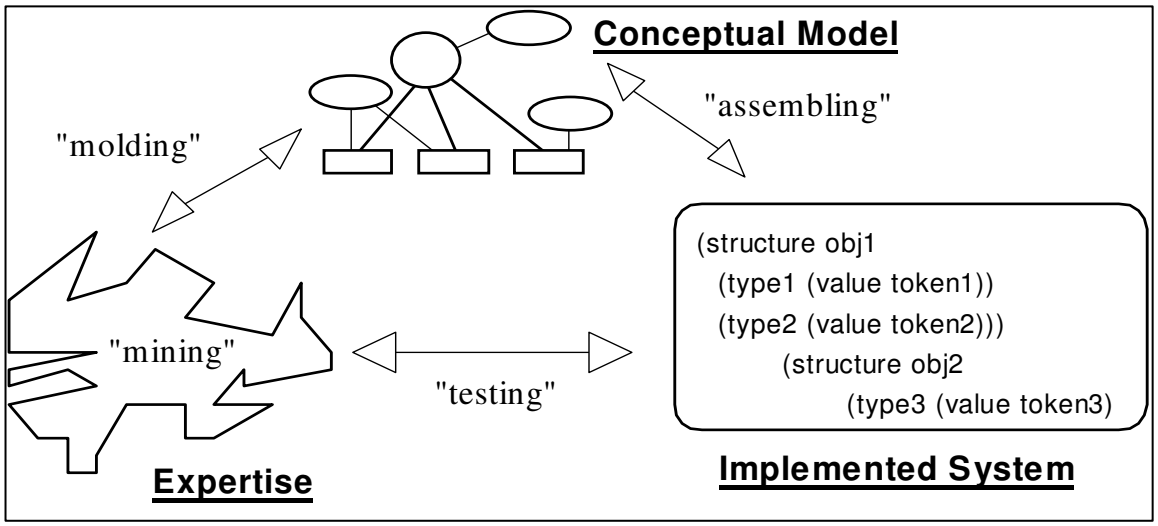


Figura 2.1. Tarefas e Resultados das Explorações

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.

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

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

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.

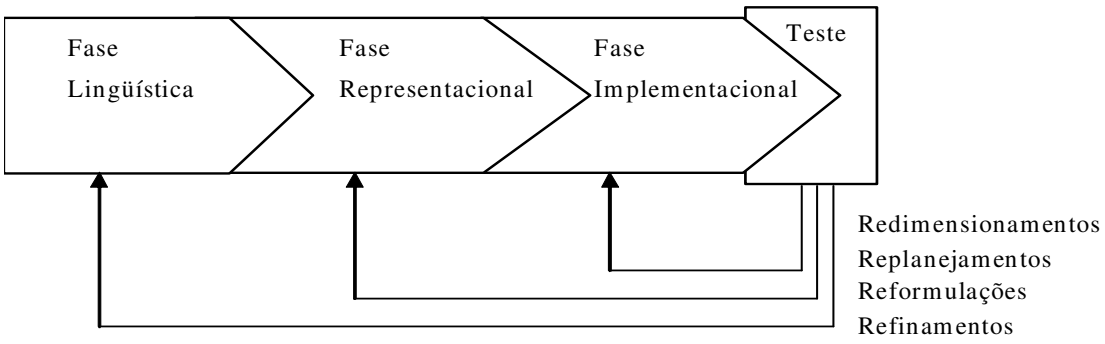


Figura 2.2. A dinâmica do processo de construção de um SPLN

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.

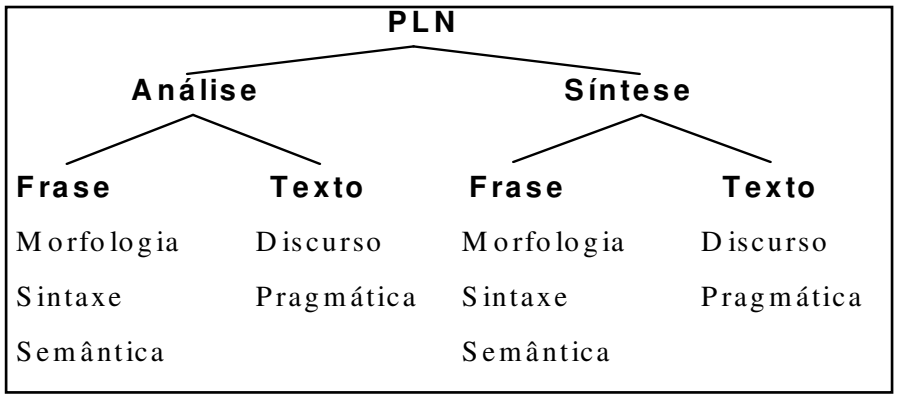


Figura 2.3. Domínios de pesquisa no âmbito do PLN

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.

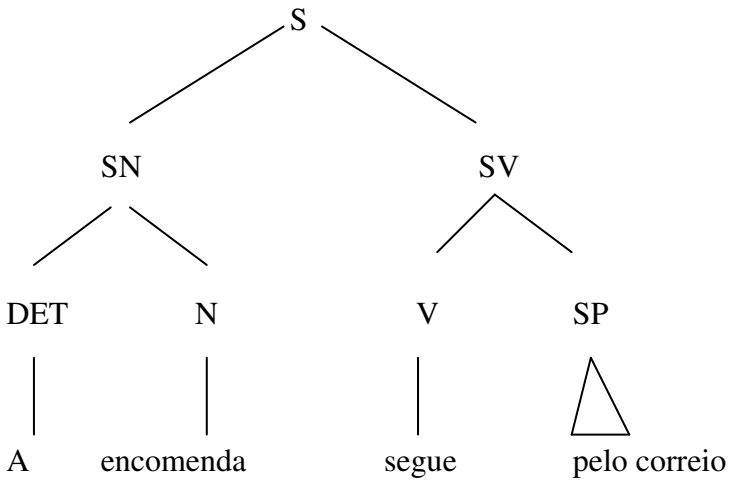


Figura 3.1. Estrutura Arbórea da sentença *A encomenda segue pelo correio*

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) Fonético-fonológicas

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.: f**e**riado; c**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) <u>Morfológicas</u>

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) <u>Sintáticas</u>

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* ⇒ sujeito, pronome pessoal [singular]
    b. *foram* ⇒ verbo [de movimento]
    c. *a padaria* ⇒ objeto indireto
    d. *hoje cedo* ⇒ 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) <u>Pragmático-discursivas</u>

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.

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.

- **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].

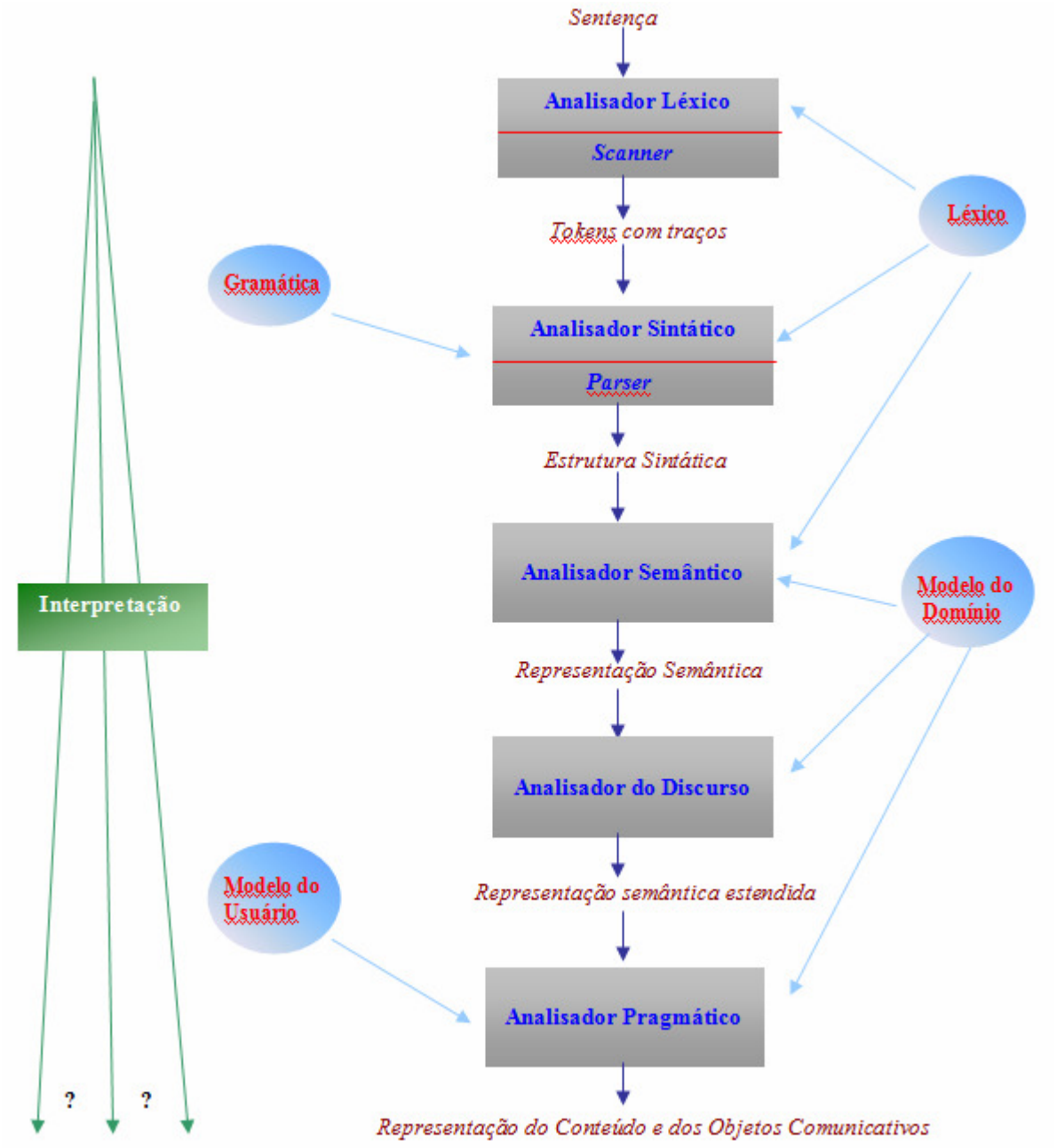


Figura 4.1. Arquitetura de um Sistema de Interpretação de Língua Natural

- **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

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

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:

  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"

[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).

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 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]

o .

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

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

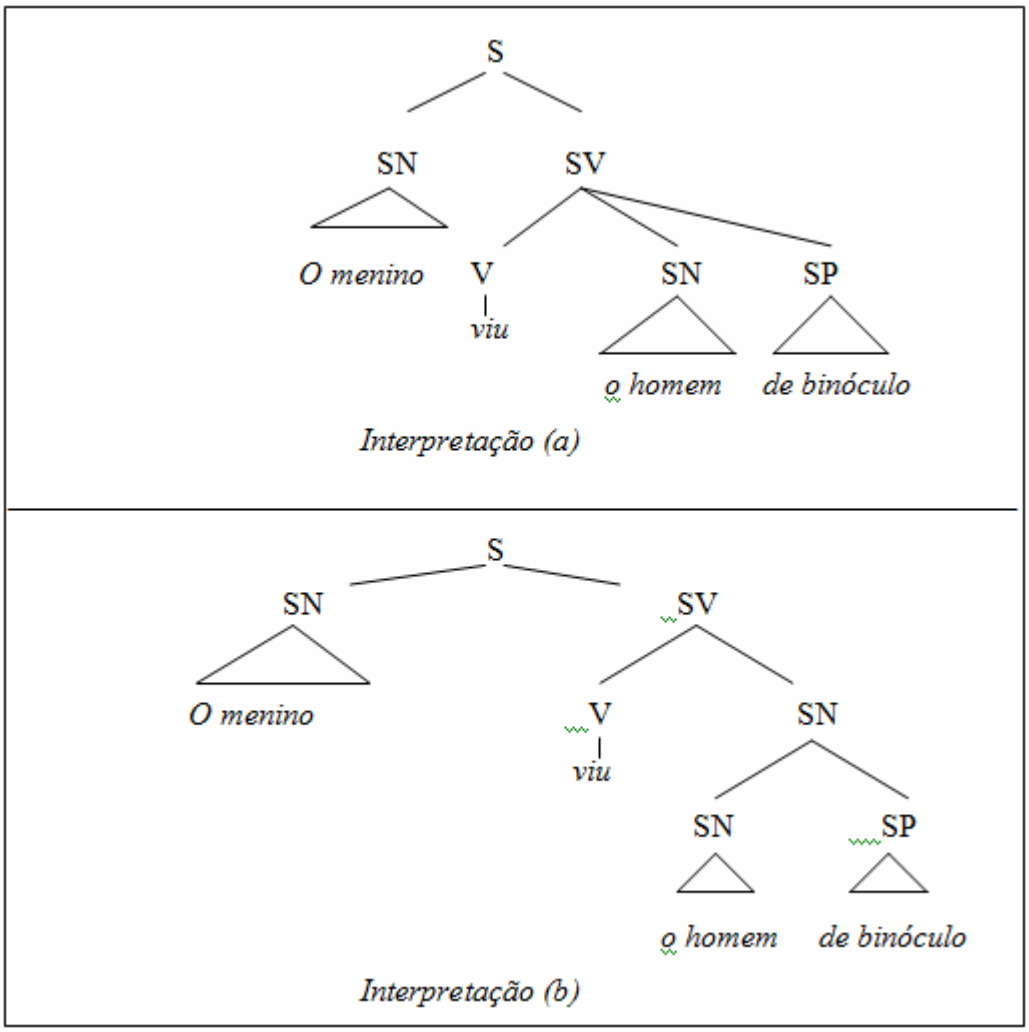


Figura 4.2. Exemplo de Ambigüidade Sintática

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

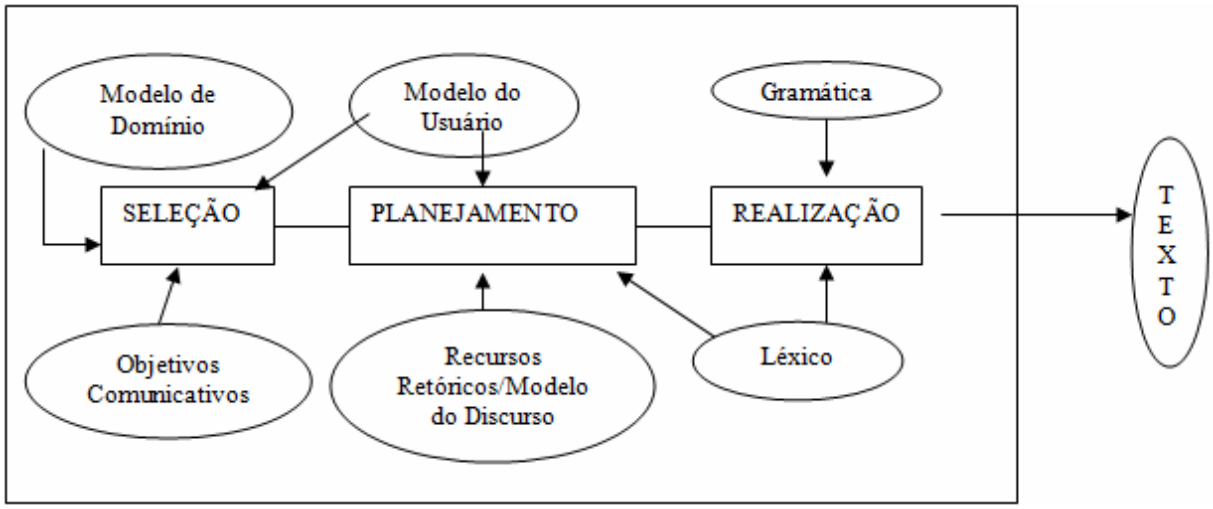


Figura 4.3. Fases Principais de um Gerador de Textos[10]

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 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), 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

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

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.

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 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 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

[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).

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 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

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

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.

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.

[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*.

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 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

[16] (Bechara, 1972)

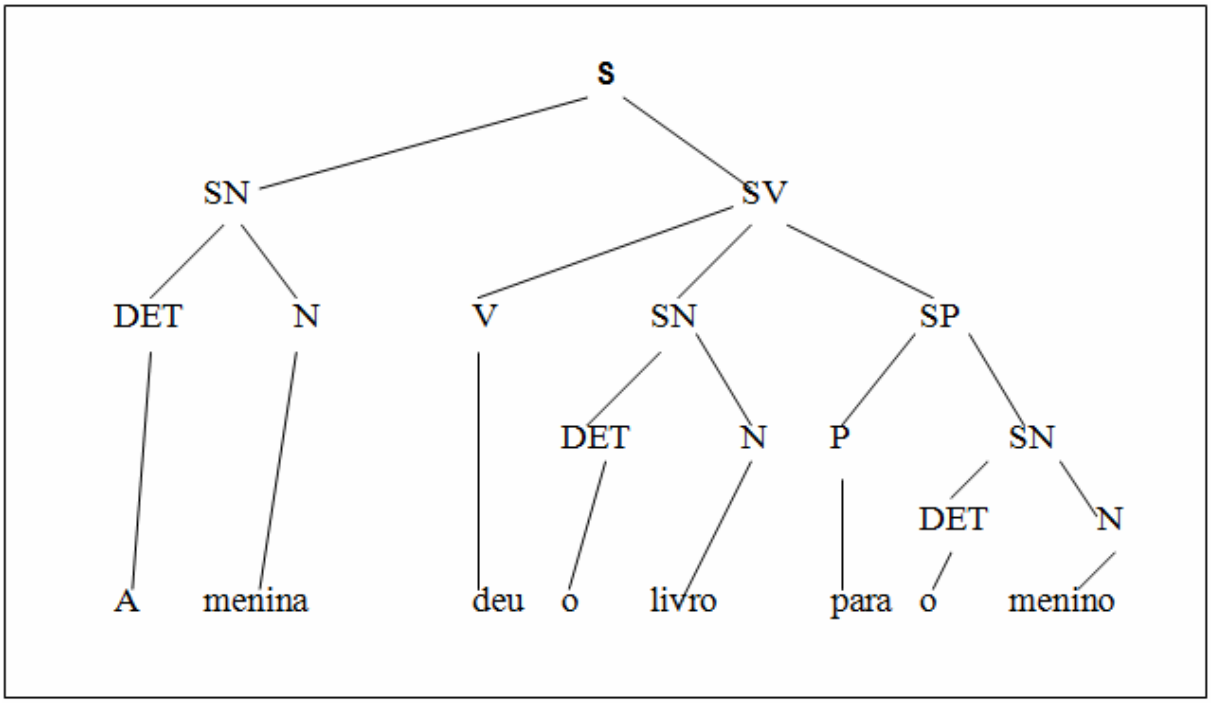
produzir enunciados por eles nunca antes ouvidos ou produzidos. De acordo com Chomsky, essa competência lingüística poderia ser expressa por um conjunto finito de regras (relacionado à *d-structure*) que, operando sobre o vocabulário da língua, regularia a produção dos enunciados lingüísticos (a *s-structure*). 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 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.

[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.

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




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

Figura 5.1. Representação e gramática para a sentença *A menina deu o livro para o menino.*[20]

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 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.

[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.

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 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 Gramma*r (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 naturais específicas, mas os princípios universais que governam a faculdade da linguagem, talvez pela insuficiência dos

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

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.

[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).

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.

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. O projeto ReGra (REvisor GRAmatical)

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. 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.

6.2. Módulo Gramatical

O primeiro passo para a elaboração do módulo lingüístico foi o levantamento de erros (ou inadequações) mais comuns entre usuários de nível médio, como secretárias e profissionais de escritório em geral, e alunos cursando o ensino médio ou ingressando a universidade. O termo "erro", aqui, refere-se ao que os gramáticos normativos consideram como forma desviante da norma culta. Como talvez pudesse ser esperado, o levantamento apontou erros de ortografia, de concordância e relacionados à crase como os mais freqüentes, seguidos de erros associados a escolhas léxicas inadequadas, principalmente por influência da oralidade. O objetivo foi o de implementar uma ferramenta voltada aos interesses de potenciais usuários. Essa escolha pressupõe a tomada de importantes decisões acerca dos itens lexicais a serem incluídos na base, a definição do que será considerado “erro” ou “inadequação”, e a elaboração de regras para detectar tais erros. Uma preocupação importante é a de minimizar o número de falsos erros (falsos positivos), ou seja, uma intervenção do sistema que pode induzir o usuário a um erro gramatical, por meio da modificação de uma estrutura lingüística originalmente correta; ou uma intervenção desnecessária do sistema que pode levar o usuário a alterar uma estrutura originalmente correta por uma outra forma correta.

Nas primeiras versões do ReGra, os erros eram detectados através de regras heurísticas implementadas na forma de redes de transição estendidas (“augmented transition networks”) (Woods, 1986), numa abordagem que se poderia chamar de "*error-driven*". O paradigma que direcionou a construção do corretor gramatical baseou-se fortemente, portanto, no estudo da língua em uso, com testes em textos reais. Para tanto, foi compilado um *corpus*, que contém textos de várias áreas do conhecimento, e inclui tanto textos já corrigidos e editados que servem como referência do uso corrente da língua escrita, quanto textos escritos por algumas classes de usuários comuns, sem correção. Pertencem à primeira classe dissertações e teses, jornais e livros. A segunda classe de textos inclui redações de vestibular e monografias, que fornecem uma amostra dos erros cometidos pelos usuários da língua.

Alguns tipos de erros podem ser detectados a partir de contextos lingüísticos bastante específicos, limitando-se à identificação, na sentença, de combinações lexicais (*patterns*) que configuram formas desviantes. Corrige-se, dessa forma, uma série de desvios, comuns para usuários inexperientes da norma-padrão da língua portuguesa escrita, como o uso indevido de crase diante de palavras masculinas e verbos, uso de ênclise nas formas do futuro, uso de ênclise e mesóclise em prejuízo de palavras atrativas, etc. Existem, também, regras de estilo, que detectam o uso de uma palavra ou expressão que não se configura como um erro gramatical, mas que é considerado impróprio para o estilo de escrita selecionado pelo usuário. Por exemplo, o uso de coloquialismos é inadequado em um texto formal, ainda que aceitável em um texto

jornalístico. O mesmo conjunto de regras (gramaticais e de estilo) pode ser aplicado a qualquer estilo. O usuário pode, durante a análise do texto, optar por desabilitar algumas regras.

O emprego de uma metodologia consistente e sistemática para a implementação de cada regra de correção foi essencial para o desenvolvimento do revisor. Foram identificadas três etapas principais:

1) identificado um tipo de erro que se deseja corrigir, é feito um estudo extensivo de gramáticas e fontes da literatura que discorram sobre o uso da língua portuguesa. Mecanismos de correção são então propostos na forma de regras para a detecção e correção desses erros.

2) as regras propostas são implementadas computacionalmente, e através de testes exaustivos com o material que compõe o corpus, são verificados falsos erros e erros não detectados. Melhorias na regra são então implementadas. Este processo é altamente interativo.

3) é dado o acabamento à regra, tanto do ponto de vista da otimização da implementação computacional, como da tomada de decisão com relação às mensagens a serem fornecidas aos usuários. As mensagens apresentam uma certa variedade, pois o sistema pode sugerir correções quando tiver certeza do erro, ou apenas alertar o usuário quanto ao uso de uma estrutura lingüística que pode ou não estar correta, dependendo do contexto.

Essas primeiras versões do ReGra apresentavam vários benefícios do ponto de vista da implementação computacional: agilidade, especificidade, rapidez, portabilidade, e disponibilidade de memória. Entretanto, seu escopo de atuação era muito limitado: problemas envolvendo itens lexicais não contíguos e estruturas recursivas não podem ser atingidos pelas estratégias heurísticas normalmente desenhadas por abordagens *error-driven*. Para prover a essas insuficiências, optou-se por analisar sintaticamente as sentenças do usuário, antes de operar a revisão propriamente dita. Isso permite aplicar regras que 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.

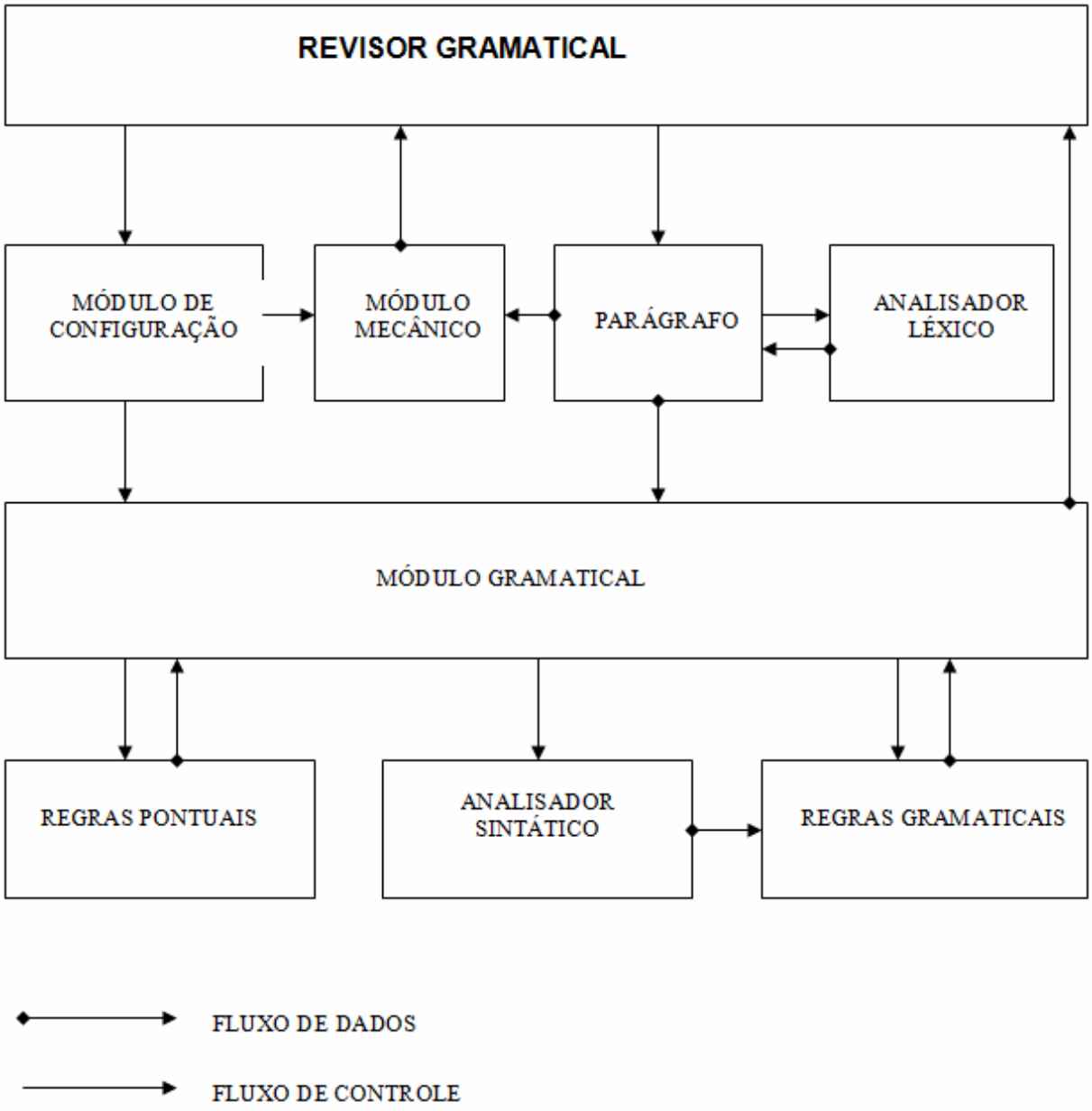


Figura 6.1. Arquitetura do ReGra

6.3. 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.

6.4. 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.

Referências da Parte I

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.

PARTE II – Introdução aos Métodos e Paradigmas de Tradução Automática

7. Introdução à TA

No cenário de globalização atual, a disseminação cada vez maior de informações multilíngües, principalmente em meio eletrônico, como a *web*, evidencia a necessidade de traduções rápidas, eficientes e baratas, para facilitar a comunicação e o compartilhamento de informações. Nesse contexto, é crescente o interesse por sistemas de tradução automática (TA), ou seja, por sistemas que permitam a tradução por computador de textos de uma língua natural para outra.

A tarefa de TA consiste em se partir de um texto-fonte, isto é, um texto escrito em uma língua natural (ou língua fonte – LF), a fim de se produzir uma versão em um texto-alvo, em uma outra língua natural (ou língua alvo – LA). Segundo Nirenburg (1987), encontrar uma forma de manter o significado, no texto-alvo, o mais próximo possível do significado do texto-fonte é o principal problema do projeto e desenvolvimento de sistemas de TA. A própria tradução humana é considerada uma das mais complexas atividades de escrita (Santos, 1998), sendo mesmo classificada como uma arte (Hutchins, 1998): a cada passo, envolve escolhas pessoais entre alternativas não codificadas. Logo, não é meramente uma questão de substituições diretas de conjuntos de símbolos, mas sim uma questão de se fazer escolhas entre valores interdependentes.

As dificuldades encontradas no desenvolvimento de sistemas de TA devem-se principalmente à necessidade de um conhecimento detalhado do texto-fonte e da situação comunicativa. Nesse contexto, o Processamento das Línguas Naturais (ou PLN) se apresenta como proposta de solução, visando satisfazer os requisitos básicos da TA, quais sejam: interpretação do texto-fonte e produção de uma representação de seu significado – com possível consideração de seu contexto ou situação discursiva – e produção do texto-alvo na LA. A evolução da pesquisa e desenvolvimento em TA foi, na verdade, fortemente influenciada pelas inovações do PLN e da lingüística formal. O PLN, como uma subárea da Inteligência Artificial, fornece uma série de técnicas computacionais para a análise e geração automática de textos em língua natural; a lingüística formal permite que a competência lingüística do falante de uma língua seja descrita em termos de um número finito de regras ou de princípios lingüísticos para gerar um número infinito de sentenças nessa língua.

As pesquisas na área de TA surgiram na década de 40 e, desde então, vários sistemas acadêmicos e comerciais vêm sendo desenvolvidos, alcançando diferentes níveis de sucesso. Esses sistemas se baseiam em diferentes métodos, a saber, TA direta, TA por transferência e TA por interlíngua. Além disso, tais sistemas podem utilizar diversos paradigmas, os quais correspondem a diferentes componentes de representação de conhecimento.

Neste relatório, apresentamos a cronologia do desenvolvimento da TA, para então identificar seus principais métodos e paradigmas. Ao final, conclusões e comentários gerais sobre esse trabalho são apresentados.

8. A evolução da tradução automática

A TA é provavelmente a aplicação mais antiga do PLN e também a primeira aplicação não numérica proposta na área da computação, na década de 40, impulsionada pelo grande número de informações disponíveis e pela idéia de que o processo computacional seria tão direto quanto a tradução humana (Nirenburg, 1987). Nessa época, os objetivos de pesquisa

eram modestos, devido a limitações de hardware e à inexistência de linguagens de programação de alto nível. A sintaxe era um tema relativamente negligenciado na área de lingüística e a semântica era geralmente ignorada. Devido a essas limitações, os sistemas de TA apresentavam resultados de baixa qualidade, exigindo um grande envolvimento humano na edição prévia e/ou posterior à execução. Nesse contexto, eram comumente considerados apenas subconjuntos de construções de uma língua natural, limitadas de acordo com regras de gramática e de vocabulário, em domínios específicos, como forma de restrição das entradas dos sistemas (Hutchins, 1998).

Com o início da guerra fria a partir de 1946, a TA passou a ser de grande interesse, principalmente para americanos e ingleses, cujo objetivo era obter informações científicas soviéticas, em geral à distância e o mais rapidamente possível. A primeira aplicação de TA nessa época foi uma calculadora científica que realizava traduções palavra por palavra, ignorando questões lingüísticas. Com ela, era possível identificar o conteúdo de um texto por uma lista de palavras-chave traduzidas, por exemplo. Em 1948, tal sistema foi refinado, para tratar desinências russas durante a análise gramatical. Já no início dos anos 50, procurou-se explorar automaticamente o contexto dos termos manipulados pela calculadora, visando solucionar problemas de ambigüidade semântica. No entanto, essa proposta era bastante equivocada: acreditava-se que os circuitos lógicos das calculadoras seriam capazes de resolver os elementos lógicos da linguagem, auxiliados pela determinação da área à qual a informação pertencia (Mateus, 1995). Diante de tal equívoco, diversos trabalhos foram desenvolvidos considerando-se a necessidade de pré-edição dos textos a serem submetidos à tradução automática.

Diretrizes mais claras para a TA foram delineadas em 1952, no congresso promovido pelo Instituto de Tecnologia de Massachusetts: deveriam ser investigadas a freqüência das palavras nos textos a serem traduzidos, as equivalências lingüísticas e outros aspectos técnicos, para só então se proceder à análise sintática e à construção, propriamente dita, dos programas de tradução correspondentes. Surgiram, assim, as abordagens fundamentais no PLN: as dirigidas por modelagem lingüística. Além disso, determinou-se, como objetivo mais próximo, o desenvolvimento de sistemas que realizassem a tradução entre duas línguas naturais em um único sentido. Considerou-se também, segundo Alfaro (1998), a possibilidade de utilizar uma língua intermediária, neutra, para se realizar a tradução, a qual viria a ser chamada posteriormente de interlíngua.

A primeira experiência de TA real, do russo para o inglês, foi realizada em 1954, na Universidade de Georgetown, com um vocabulário reduzido (250 palavras), textos cuidadosamente selecionados e 6 regras de sintaxe. Essa experiência foi considerada satisfatória. Segundo Hutchins (1998), a partir desse resultado, órgãos que patrocinavam projetos de TA passaram a acreditar que poderiam ser desenvolvidos sistemas que produzissem traduções de boa qualidade em poucos anos. Entretanto, tal nível de qualidade ainda era dependente da evolução de hardware, do surgimento ou refinamento das linguagens de programação de alto nível existentes e, principalmente, do desenvolvimento das pesquisas para a análise sintática, sobretudo referentes à exploração de gramáticas formais, dentre as quais o grande marco, na época, foi a gramática normativa de Chomsky (1957).

A partir de então, as pesquisas em TA passaram a considerar como objetivo o desenvolvimento de sistemas completamente automatizados produzindo traduções de alta qualidade em domínios amplos. A ênfase nas pesquisas tornou-se a busca por teorias e métodos que permitissem alcançar tais objetivos.

No final dos anos 50, além dos americanos, outros países europeus começaram a explorar e investir na TA. Buscava-se ainda transformar os estudos lingüísticos em uma ciência exata, empregando-se métodos matemáticos. No entanto, os primeiros projetos de TA resultantes desses investimentos não alcançaram suas ambiciosas metas. O progresso foi

muito mais lento do que se esperava, devido à complexidade de tratamento computacional dos aspectos formais, teóricos, da lingüística e aos aspectos da própria TA. A lingüística formal não conseguia explicar, por exemplo, os problemas estruturais, funcionais e práticos da TA. Como resultado, houve um descrédito generalizado na TA, culminando com um relatório do ALPAC (*Automatic Language Processing Advisory Committee*) – comitê composto pelos patrocinadores americanos – em 1966, declarando que a TA havia falhado em atingir suas metas, uma vez que não existia nenhum sistema completamente automático capaz de produzir traduções de boa qualidade. Esse relatório também era fortemente negativo com relação às chances futuras de sucesso da TA, o que provocou um corte radical de verbas governamentais norte-americanas. Alguns poucos projetos foram mantidos, agrupados, sobretudo, em três classes: ferramentas computacionais para auxílio à tradução humana, sistemas de TA envolvendo a interação humana e pesquisas teóricas sobre melhorias dos métodos de TA (Hutchins, 1998).

Nirenburg (1987) diz que é importante lembrar que os esforços iniciais tiveram grande importância para o estudo das línguas naturais e do seu processamento via computador, pois contribuíram para o desenvolvimento de várias áreas, dentre as quais destacam-se a Lingüística Moderna, a Lingüística Computacional e a própria Inteligência Artificial. Após o período "negro" do PLN, alguns fatos inovadores reativaram o interesse pela TA no início da década de 80: foi criada a Comunidade Econômica Européia; houve uma explosão da informatização, com grandes avanços de técnicas de computação e da inteligência artificial; as pesquisas e o desenvolvimento de novas teorias no âmbito da lingüística formal (em especial, as teorias de Chomsky) possibilitaram o aprofundamento das investigações no campo da semântica e o processamento automático de várias línguas naturais com base em gramáticas de análise e de geração. Além disso, a TA se enquadrou em um contexto mais realista, no qual aceitava-se que, mesmo imperfeita, ela poderia ser muito útil.

As reflexões sobre as reais possibilidades da TA originaram novas metas e interesses, caracterizando sistemas de diversas naturezas. Sistemas de recuperação de informação, por exemplo, em um ambiente de TA, podem ser úteis mesmo que suas traduções não sejam muito boas. Basta que permitam a compreensão das idéias principais do texto sob exploração (Slocum, 1985). Com foco em situações e objetivos específicos, a TA passou a receber apoio governamental maciço, principalmente na Europa e Japão, no final da década de 80. Acreditava-se em idéias como a de Slocum (1985), de que um sistema de TA de boa qualidade é aquele que gera um texto que permite uma revisão sem grandes problemas e cuja operação completa (incluindo essa revisão) oferece uma boa relação custo-benefício. Como conseqüência, mantiveram-se as áreas de investigação e desenvolvimento de aplicativos computacionais de auxílio à tradução, além de programas de TA prevendo a intervenção humana. Alguns produtos passaram a ser desenvolvidos, como o *Systran* (http://www.systransoft.com) e o *Eurotra* (http://www.ccl.kuleuven.ac.be/about/EUROTRA.html), sistemas americano e europeu, respectivamente, em constante desenvolvimento.

A partir do final dos anos 80, vários fatores contribuíram para a definição de um novo cenário de TA (Hutchins, 1998): 1) diversos sistemas de tradução comerciais foram disponibilizados no mercado, para amplo uso por tradutores humanos profissionais ou por usuários comuns; 2) cresceu significativamente a aquisição de computadores de uso pessoal, com a perspectiva do aumento de ferramentas de comunicação dedicadas; 3) teve início o desenvolvimento de sistemas particulares, de domínio específico; 4) o crescimento das redes de telecomunicação envolvendo muitas línguas conduziu à demanda de dispositivos de tradução, em tempo real, de grandes volumes de dados eletrônicos; 5) a grande disponibilidade de bancos de dados e recursos de informações em muitas línguas diferentes

levou à necessidade de dispositivos de busca e acesso que incorporassem módulos de tradução.

Nesse cenário, o problema computacional da TA foi praticamente superado: diversos serviços de TA são oferecidos na Internet; grupos de pesquisa em TA permitem que o público realize testes *on-line* dos programas em desenvolvimento; os dicionários e as gramáticas necessários para o funcionamento de sistemas de TA podem ser ampliados pelos próprios usuários; a velocidade e a eficiência das consultas aos bancos de dados são cada vez maiores. No entanto, restam importantes questões de cunho lingüístico a resolver (semântico e pragmático-discursivo, principalmente), tais como ambigüidades, referências anafóricas, etc. Como conseqüência, o desenvolvimento de sistemas completamente automatizados, que consideram questões lingüísticas e extralingüísticas de forma profunda, principalmente em domínios abertos ou línguas naturais irrestritas, após mais de 50 anos de pesquisa, ainda é um desafio para a área de TA.

De fato, segundo Kay (1994), ainda hoje alcançam resultados mais práticos e significativos os sistemas de TA desenvolvidos em contextos limitados, como é o caso do *Taum-Meteo* (Isabelle, 1987), utilizado para produzir boletins meteorológicos bilíngües, do inglês para o francês, cuja linguagem é altamente estilizada, regular e específica. Porém, sistemas baseados em sublínguas não constituem interesse para a tradução entre falantes de duas línguas naturais, por serem altamente restritos pela comunidade de uso.

Em domínios abertos, por outro lado, geralmente os textos traduzidos são compreensíveis, mas nem sempre gramaticais e raramente fluentes, implicando a necessidade de revisão humana na fase de pós-processamento.

Alguns sistemas de TA servem de auxiliares para tradutores humanos, no sentido de que realizam uma pré-tradução do texto, a ser editada/refinada pelos tradutores humanos, a exemplo dos tradutores *Trados Workbench* (http://www.trados.com/), *IBM Translation Manager* (http://www-4.ibm.com/software/ad/translat/) e *Déjavu* (http://www.atril.com). Outros, ainda, consideram a pré-edição do documento original, de modo a apresentá-lo em uma linguagem mais simples, como a usada pela Xerox no *Systran* (http://www.systransoft.com), criado inicialmente para traduzir seus manuais técnicos em várias línguas.

Sistemas de TA que consideram alguma forma de edição humana, seja ela feita previamente, durante a tradução ou posteriormente, são chamados de *Human-Aided Machine Translation* ou, simplesmente, HAMT. Quando servem de auxílio à tradução humana, são chamados *Machine-Aided Human Translation* ou, simplesmente, MAHT. Esses últimos incluem ferramentas de acesso a dicionários e enciclopédias, recursos de processamento de textos, verificação ortográfica e gramatical, entre outras (Boitet, 1994).

Atualmente, a *web* certamente é responsável pelo novo incentivo à TA. Com a popularização da Internet, cresceram consideravelmente a oferta e a procura de programas de TA. Diversos sistemas são capazes de traduzir páginas da Internet *on-line*, mensagens de correio eletrônico ou conversas via programas de *chat*.

É nesse contexto de apelo à necessidade de tradução de textos disponíveis de forma eletrônica que se fundamenta o projeto e desenvolvimento de parte significativa dos sistemas de TA atuais, com base em diferentes métodos de tradução. Consideram-se três métodos, descritos a seguir: tradução direta, por transferência e por interlíngua.

9. Métodos de TA

De acordo com Dorr et al. (2000), dois tipos de informação podem ser utilizados para classificar um sistema de TA: seu método e seu paradigma. Os **métodos** referem-se ao projeto de processamento, ou seja, à organização global do processamento e de seus vários módulos,

enquanto os **paradigmas** referem-se aos componentes de representação de conhecimento que auxiliam o projeto de processamento global.

Há três diferentes métodos de TA: **TA direta**, **TA por transferência** e **TA por interlíngua**. Esses métodos podem ser agrupados em duas categorias: a **tradução direta** e a **tradução indireta**, esta incluindo os dois últimos métodos.

A Figura 1 (Dorr et al., 2000, p. 13) delineia os diversos níveis de profundidade do conhecimento a ser manipulado por cada um dos métodos. É importante notar que o mesmo conhecimento lingüístico pode ser utilizado por diferentes métodos (por exemplo, o conhecimento semântico pode ser utilizado tanto no método por interlíngua quanto no método por transferência).

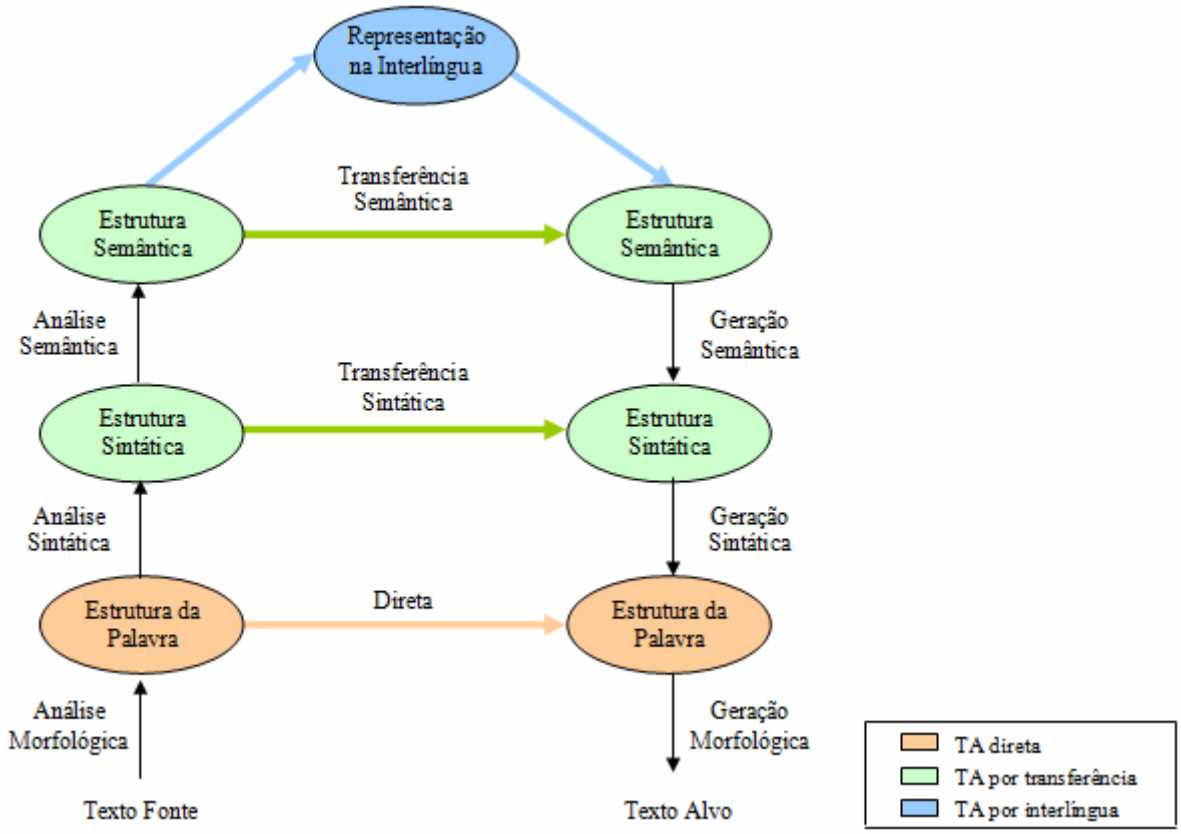


Figura 1. Níveis de profundidade do conhecimento nos sistemas de TA.

A seguir, são descritas as principais características de cada método.

9.1. Método direto

A TA direta transforma as sentenças da LF em sentenças da LA sem utilizar representações intermediárias, procurando realizar o mínimo de processamento lingüístico possível. Esse processamento pode variar, incluindo a simples substituição das palavras de uma sentença-fonte por sua(s) correspondente(s) na LA (tradução palavra-por-palavra) ou a realização de tarefas mais complexas, como a reordenação das palavras na sentença-alvo e a inclusão de preposições.

Geralmente, o processo de tradução compreende a análise sintática simplificada, a substituição das palavras-fonte por suas equivalentes na LA, utilizando, para tanto, um

dicionário bilíngüe, e a reordenação das palavras de acordo com as regras da LA, a partir de informações sintáticas de ambas as línguas (Arnold et al., 1993). A Figura 2 ilustra esse processo.

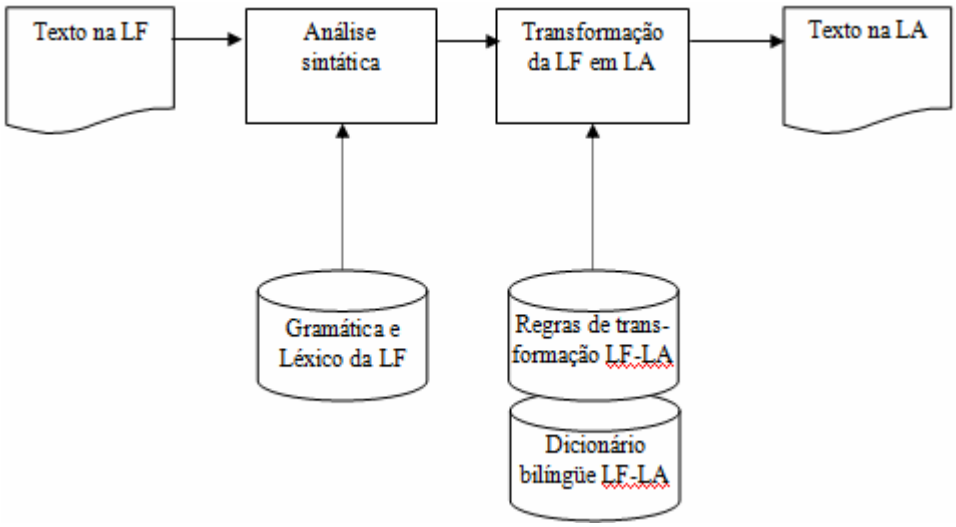


Figura 2. TA pelo método direto.

Sistemas de TA direta são comumente construídos para um único par de línguas e a definição do número de estágios depende, na verdade, além do nível de qualidade pretendido, da proximidade das línguas envolvidas. Ward (1999) cita um exemplo de um sistema de tradução do japonês para o inglês consistindo de seis estágios:

1) Análise lexical e morfológica, na qual a sentença é separada em palavras;
2) Transferência lexical das palavras na sentença, por meio do dicionário bilíngüe;
3) Adequações relacionadas às preposições (a resolução preposicional do japonês é significativamente distinta da resolução para o inglês, merecendo um módulo isolado);
4) Transformações das estruturas sentenciais. Por exemplo, a estrutura dos constituintes da forma SOV (Sujeito Objeto Verbo), utilizada na língua japonesa, é transformada na estrutura SVO (Sujeito Verbo Objeto), utilizada na língua inglesa;
5) Tarefas diversas, como a inserção de artigos;
6) Geração morfológica, na qual as palavras em inglês são flexionadas.

No que se refere aos resultados esperados por sistemas de TA direta, pode-se identificar as seguintes características (Arnold et al., 1993):

1) O sistema pode ser considerado robusto, pois sempre apresenta algum resultado, mesmo que seja ruim, por conta de problemas como a não tradução de algumas palavras (por não existirem no dicionário), ou a geração de construções gramaticais desconhecidas (por não existirem regras de transformação adequadas).
2) Não há como garantir que a sentença traduzida seja realmente uma sentença gramatical na LA – o resultado pode ser um inelegível emaranhado de palavras.
3) A qualidade do resultado dos sistemas que fazem somente a correspondência direta entre palavras tende a ser muito baixa, o que justifica a exploração de sistemas de tradução direta mais complexos.

Segundo Dorr et al. (2000), de um modo geral, as traduções são pobres; no entanto, se limitadas a domínios restritos e a textos simples, podem ser bastante úteis, principalmente para especialistas naquele domínio.

Quanto à utilização desse método, grande parte dos sistemas de TA comerciais, principalmente os mais antigos, foi desenvolvida com base nele, a exemplo do *Systran* (http://www.systransoft.com), cujo processo de tradução consiste basicamente de buscas em um dicionário, palavra a palavra.

9.2. Método indireto

Nos sistemas de tradução indireta, a análise da LF e a geração da LA constituem processos independentes, cada qual tratando somente dos problemas da língua envolvida. Diferentemente do método direto, esses sistemas se baseiam na idéia de que a TA de alta qualidade requer conhecimento lingüístico (e eventualmente extralingüístico) de ambas as línguas, assim como das diferenças entre elas. Esse conhecimento é representado por linguagens intermediárias entre as línguas fonte e alvo. Assim, as sentenças fonte são primeiramente transformadas numa representação na linguagem intermediária e, a partir dela, são geradas as sentenças alvo.

Existem dois métodos de TA indireta: por transferência e por interlíngua. Dependendo do método, a representação intermediária pode ser única, independente de língua (tradução por interlíngua), ou dependente de língua (tradução por transferência). Neste último caso, são necessárias duas linguagens de representação intermediária: uma para a LF e outra para a LA.

Para codificar o conhecimento das línguas fonte e alvo e o conhecimento das relações entre elas, basicamente são necessários os seguintes componentes:

1) gramáticas e léxicos substanciais de ambas as línguas, os quais são utilizados tanto na análise das sentenças fonte, quanto na geração das sentenças alvo;
2) dicionários bilíngües para as regras de substituição de palavras;
3) no caso da tradução por transferência, uma gramática comparativa, ou seja, um conjunto de regras de transformação para relacionar a representação intermediária da LF com a representação intermediária da LA;
4) no caso de tradução por interlíngua, um conjunto de regras de transformação para relacionar a interlíngua com as línguas fonte e alvo.

Dependendo do método de TA indireta e também do nível de profundidade da análise realizada, outros componentes podem ser necessários, conforme descrito a seguir.

9.2.1. TA por transferência

Na TA por transferência, a tradução consiste nos seguintes passos (Figura 3): 1) alteração da estrutura e palavras da sentença de entrada resultando em uma representação intermediária da LF (fase de análise); 2) transformação dessa representação em uma estrutura intermediária da LA (fase de transferência); e 3) geração da sentença na LA (fase de geração), a partir dessa estrutura.

A fase de análise pode envolver processos complexos como as análises semântica e pragmática, mas, em geral, são mais comuns sistemas que se limitam à análise sintática, gerando como representação intermediária uma estrutura de árvore. Nesse caso, a fase de transferência converte essa estrutura da LF em uma estrutura de árvore da LA, por meio de regras de mapeamento entre as duas línguas naturais, que indicam as correspondências lexicais e sintáticas entre tais estruturas. Para tanto, é necessário representar o conhecimento contrastivo (i.e., comparativo) das duas línguas, o qual envolve a especificação de suas diferenças normativas e lexicais. A fase de geração transforma a estrutura de árvore da LA na sentença final, propriamente dita, utilizando a gramática e o léxico da LA.

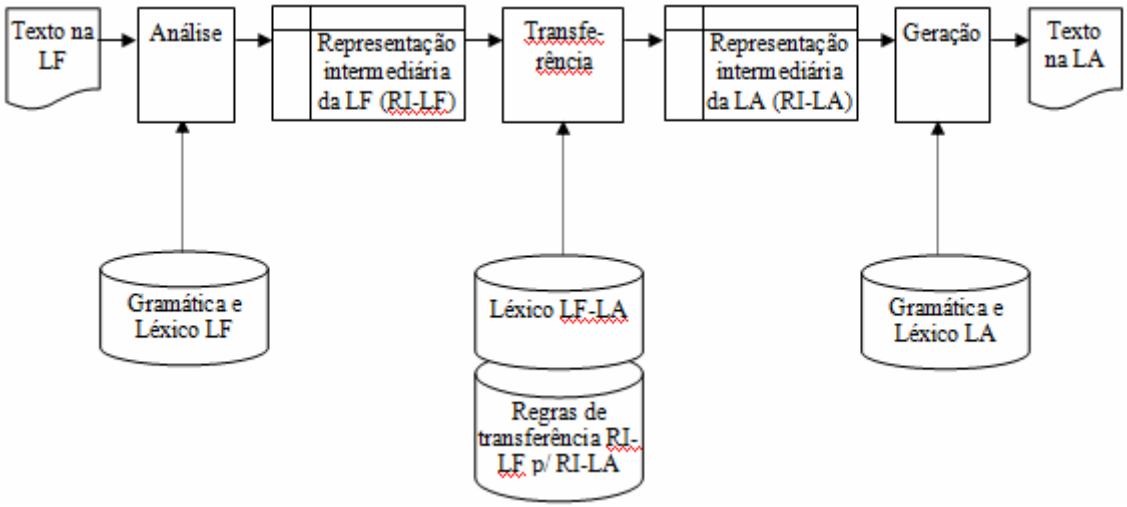


Figura 3. TA pelo método indireto por transferência

Da mesma forma que na tradução direta, a profundidade do módulo de análise depende da proximidade das línguas: quanto mais próximas forem as línguas, mais superficial pode ser a análise. Porém, para traduções de alta qualidade, mesmo quando as línguas são próximas, análises mais profundas que a sintática são necessárias. Assim, diversos sistemas incorporam aos seus processos também informações semânticas, que podem ser representadas, por exemplo, por *frames*, isto é, coleções de atributos e valores. Nesse caso, um módulo de análise semântica é responsável por preencher os atributos semânticos de componentes sentenciais, por exemplo, a partir de uma árvore sintática; o módulo de transferência mapeia, então, esse *frame* em um outro *frame* da LA, o qual é convertido para a sentença na LA.

Análises ainda mais profundas, baseadas em informações de contexto (pragmático-discursivas) dificilmente são realizadas em sistemas de TA por transferência, dada sua grande complexidade. Na verdade, até mesmo as informações de ordem semântica são geralmente incorporadas somente para resolver problemas limitados.

Segundo Dorr et al. (2000), a qualidade global dos sistemas de transferência sintática, é maior que a dos sistemas de tradução direta, mas tende a ser menor que a dos sistemas que empregam uma análise mais profunda do texto fonte, como aqueles que utilizam o método por interlíngua.

Alguns exemplos de abordagens de TA por transferência são o projeto *Eurotra* (http://www.ccl.kuleuven.ac.be/about/EUROTRA.html), cujo objetivo é a criação de um ambiente multilingüe para todas as línguas da Comunidade Européia; e o *Vermobil* (Wahlster, 1993), patrocinado pelo governo da Alemanha, que reconhece textos falados em alemão e os traduz para textos falados em inglês.

9.2.2. TA por interlíngua

Devido à dificuldade de se estabelecer regras de transferência e recursos lingüísticos comparativos (como gramáticas) efetivos, necessários aos sistemas desenvolvidos sob o método de TA por transferência, e também à complexidade inerente a esses sistemas, houve o interesse pela definição de um nível de análise tão profundo a ponto de permitir descartar os componentes contrastivos entre as línguas em foco, presentes na tradução por transferência. O objetivo era fazer com que a saída da análise da LF correspondesse diretamente à entrada do componente de geração na LA. Representações nesse nível deveriam capturar, assim, o significado a ser transmitido, independentemente da língua natural em questão. Esta é

justamente a função da **interlíngua**: permitir extrair a representação do significado da sentença fonte para, a partir dela, gerar a sentença na LA.

Nesse cenário de TA por interlíngua, o processo de tradução é feito de acordo com as seguintes etapas (Figura 4): 1) análise completa do texto na LF, extraindo seu significado e representando-o na interlíngua; e 2) geração do texto na LA, partindo da representação interlingual e expressando o mesmo significado. Nesse contexto, o processo de geração do texto na LA caracteriza-se mais como uma paráfrase que como uma tradução, podendo ser perdidos o estilo e o foco do texto original.

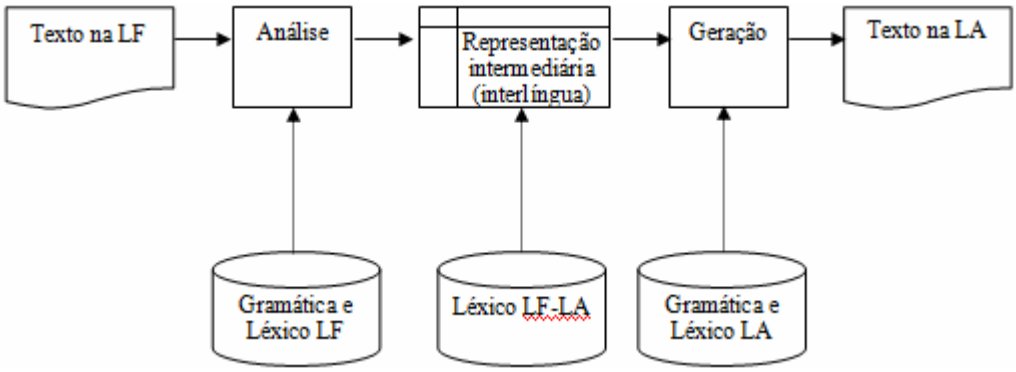


Figura 4. TA pelo método indireto por interlíngua

Uma das maiores dificuldades da TA por interlíngua é a própria especificação da interlíngua, a qual deve ser independente de qualquer língua natural, para representar o significado de suas sentenças de modo uniforme e consistente. Os principais problemas são: 1) como escolher o léxico dessa interlíngua, que deve ser composto por conceitos primitivos que permitam expressar o significado de todo o vocabulário das línguas em questão e 2) como definir a gramática da interlíngua, a qual deve conter regras suficientemente robustas para relacionar todos os possíveis conceitos primitivos.

De acordo com Ward (1999), é preciso realizar análises exaustivas sobre a semântica do domínio, de modo a formalizar os tipos de entidades que existem e o seu relacionamento. Isso pode ser facilitado a partir da definição de uma **ontologia**, isto é, de um modelo de mundo sobre um dado domínio, contendo todos os conceitos (palavras) que podem ser expressos naquele domínio e suas relações (Uschold e Gruninger, 1996).

As principais vantagens da TA por interlíngua são as seguintes:

1) A facilidade com que os sistemas podem ser estendidos, pois novas línguas podem ser adicionadas a um custo relativamente baixo; e

2) A possibilidade de incluir níveis de representação mais profundos, como o pragmático-discursivo, resultando em sistemas com potencial maior qualidade que os desenvolvidos sem esses níveis.

Uma vez que sistemas de tradução por interlíngua requerem um conhecimento extensivo, seu desempenho depende, em grande parte, da coleção e representação eficiente de grandes quantidades de conhecimento sobre o domínio em questão. Isso deve ser feito por meio de teorias ou linguagens de representação (semânticas e/ou pragmático-discursivas), como, por exemplo, a Teoria Dependência Conceitual (Schank, 1975), a Teoria de Casos (Fillmore, 1968), a Semântica Conceitual (Jackendoff, 1990) e a UNL (UNL, 2001).

Em vista das vantagens oferecidas pela TA por interlíngua, nos últimos anos vários sistemas baseados nesse método vêm sendo desenvolvidos, envolvendo diferentes línguas e objetivos específicos. O foco das pesquisas consiste, geralmente, na especificação de uma interlíngua que seja adequada para todas as línguas envolvidas, além de uma análise

conceitual que permita produzir traduções de qualidade aceitável (Dorr et al., 2000). A seguir, são descritos brevemente alguns desses sistemas.

Exemplos de sistemas de TA por interlíngua

• **TRANSLATOR** (Nirenburg et al., 1987): sistema que explora o paradigma KBMT, utilizando, além dos módulos de análise e geração dos sistemas de TA por interlíngua, um módulo de *enriquecimento*, o qual dispõe de uma base de conhecimento (um dicionário e uma gramática da interlíngua) para, a partir do texto na interlíngua gerado pelo analisador, realizar algumas inferências de modo a acrescentar a esse texto informações sobre o domínio em questão. A estrutura enriquecida é então repassada ao gerador da LA. O TRANSLATOR possui duas linguagens de representação do conhecimento: uma para descrever os conceitos da interlíngua no dicionário (DRL) e uma para descrever a sintaxe da interlíngua na gramática (GRL).

• **KBMT-89** (Nirenburg e Goodman, 1991): sistema para traduções entre inglês e japonês, baseado em conhecimento (KBMT), tendo como domínio manuais de manutenção e instalação de computadores pessoais. A representação da interlíngua é feita por *frames*. O sistema se baseia no conhecimento expresso por uma ontologia muito bem definida, criada a partir do formalismo *Ontos* (Carlson e Nirenburg, 1992).

• **ULTRA** – *Universal Language TRAnslator* (Farwell e Wilks, 1991): sistema de TA multilingüe que realiza traduções de palavras individuais ou de sentenças entre as línguas chinesa, inglesa, alemã, espanhola e japonesa. Contempla sentenças declarativas, interrogativas e imperativas e construções conjuntivas, além de tratar de algumas desambigüizações de sentido e dependências contextuais (anafóricas e elípticas, por exemplo).

• **KANT** (Mitamura et al., 1991; Nyberg e Mitamura, 1992): sistema de TA multilingüe, baseado no paradigma KBMT, que traduz documentos técnicos do inglês para o japonês, francês e alemão, com qualidade elevada. O domínio das traduções é específico (manuais técnicos) e a entrada para o sistema é na forma de textos em uma linguagem simplificada (isto é, os textos são pré-processados). As principais características da arquitetura do sistema são a aquisição de conhecimento semi-automatizada e a interpretação semântica profunda.

• **UNITRAN** (Dorr, 1992): sistema baseado na semântica lexical, desenvolvido para traduções bidirecionais entre espanhol, inglês e alemão. A representação interlingual adotada é a LCS (*Lexical Conceptual Structure*), uma extensão da estrutura conceitual de Jackendoff (1990), escolhida principalmente por prover um mapeamento sistemático entre a interlíngua e as estruturas sintáticas.

• **UNL** – *Universal Networking Language* (Uchida et al., 1999; UNL, 2001): projeto que contempla a TA multilingüe, incluindo atualmente 14 línguas, envolvendo o trabalho cooperativo de diversos países. Para esse projeto, a UNU (Universidade das Nações Unidas) criou e disponibilizou para todos os grupos de projeto e desenvolvimento UNL os seguintes recursos: 1) a especificação da interlíngua UNL; 2) codificadores (da LF para a IL) e decodificadores (da IL para a LA) genéricos a

serem customizados para cada língua natural em foco; e 3) uma base de conhecimento para representar informações universais a respeito de todos os conceitos do repositório da UNL, inclusive informações ontológicas, na forma de uma hierarquia conceitual.

9.3. A complexidade dos métodos de TA

A comparação da complexidade dos três métodos de TA apresentada aqui se dá em função do número de módulos necessários ao processo de tradução. Considera-se que os sistemas de TA podem trabalhar com tradução **unidirecional**, isto é, para cada par de línguas, uma língua é fonte ou alvo, mas não as duas coisas; ou **bidirecional**, isto é, para cada par de línguas L1 e L2, a tradução pode ocorrer tanto de L1 para L2, quanto de L2 para L1. Em ambos os casos, assume-se que tais módulos realizam análise/transferência/geração em apenas um sentido, por exemplo, um módulo para transferência entre representações intermediárias do português para o inglês não realiza a transferência entre representações intermediárias do inglês para o português.

No método direto, considerando apenas o módulo de transformação da LF em LA, para cada par de línguas é necessário um módulo de transferência, se a tradução for unidirecional. Assim, para *n* línguas, são necessários *n-1* módulos. Se a tradução for bidirecional, são necessários dois módulos de transferência, totalizando *n*(n-1)* módulos.

O método por transferência, para *n* línguas, na tradução unidirecional, envolve *n-1* módulos de transferência, além de *1* módulo de análise e *n-1* módulos de geração, totalizando *2n-1* módulos. Já para a tradução bidirecional são necessários *n*(n-1)* módulos de transferência, *n* módulos de análise, e *n* módulos de geração, somando *n*(n+1)* módulos.

No método por interlíngua, a quantidade de módulos necessários é proporcional ao número de línguas que o sistema manipula, e não ao quadrado desse número, como no método por transferência. Para cada nova língua, são criados apenas os módulos de análise e/ou geração, ou seja, as regras de mapeamento da nova língua para a interlíngua e/ou vice-versa. Assim, para *n* línguas, são necessários *n* módulos para a tradução unidirecional e *2n*, para a tradução bidirecional.

De modo sintetizado, a Tabela 1 apresenta as características de complexidade dos três métodos de tradução para *n* línguas, considerando sistemas de tradução uni ou bidirecional.

Tabela 1. Complexidade dos métodos de TA

Método	**Nº de módulos (para tradução unidirecional)**	**Nº de módulos (para tradução bidirecional)**
Direto	n-1	n*(n-1)
Transferência	n^2	n*(n+1)
Interlíngua	n	2n

10. Paradigmas de TA

Sistemas desenvolvidos de acordo com os diferentes métodos descritos podem se basear em conhecimento profundo, fundamental ou lingüístico – **paradigma fundamental** – ou em conhecimento superficial ou empírico – **paradigma empírico** (Arnold et al., 1993). Esta seção descreve, brevemente, os modelos de representação do conhecimento fundamental e algumas abordagens empíricas atuais.

As combinações possíveis entre métodos e paradigmas são várias. O uso de um paradigma nem sempre exclui o uso de outros, pelo contrário, muitos sistemas de TA são baseados em abordagens híbridas, as quais podem incluir também combinações entre diferentes métodos, cada método sendo responsável pelo tratamento de determinados aspectos da tradução (sistemas *multi-engine*).

10.1. Paradigmas fundamentais

Os modelos de TA fundamentais são aqueles que empregam teorias lingüísticas bem definidas, utilizando restrições sintáticas, lexicais ou semânticas, sobre as línguas naturais envolvidas. A seguir, são descritos alguns dos diferentes tipos de conhecimento que caracterizam esses modelos.

10.1.1. TA baseada em regras

Sistemas de TA baseados em regras (*Rule-Based Machine Translation*, ou RBMT) são caracterizados por representar o conhecimento por meio de regras de diferentes níveis lingüísticos, para a tradução entre as línguas fonte e alvo. Por exemplo, para a transferência lexical, as características e restrições de itens lexicais individuais são codificadas num mecanismo de controle, por meio de regras, e não no léxico.

Rosseta (1994) descreve um exemplo de sistema RBMT interlingual, dividindo as regras de tradução em duas categorias: 1) regras que fazem o mapeamento de árvores sintáticas em estruturas de significado; e 2) regras que fazem o mapeamento de itens lexicais em árvores sintáticas.

10.1.2. TA baseada em conhecimento

O paradigma baseado em conhecimento (*Knowledge-Based Machine Translation*, ou KBMT) define sistemas baseados em regras que utilizam conhecimento profundo, lingüístico ou extralingüístico, de um domínio, permitindo que o sistema possa tecer inferências sobre os conceitos manipulados. Segundo Kay (1994), a maior justificativa para utilização de sistemas KBMT é que a tradução depende fortemente de informações e características extralingüísticas, de senso comum e de conhecimento do mundo. A representação do conhecimento pode envolver o desenvolvimento de ontologias e modelos de domínio. Alguns exemplos de sistemas dessa categoria são o *Translator* (Nirenburg et al., 1987), o KBMT-89 (Nirenburg e Goodman, 1991), o KANT (Mitamura et al., 1991; Nyberg e Mitamura, 1992) e o Projeto UNL (Uchida, 1999; UNL, 2001), os quais foram brevemente apresentados.

10.1.3. TA baseada em léxico

Sistemas baseados em léxico (*Lexicon-Based Machine Translation*, ou LBMT) são aqueles que fornecem regras para relacionar as entradas lexicais de uma língua às entradas lexicais de

outra língua. Um exemplo de sistema dessa categoria é o LTAG (Abeillé et al., 1990)[23], para traduções do inglês para o francês e vice-versa. O LTAG é um sistema de transferência que utiliza TAGs – *Tree Adjoining Grammars* (Joshi, 1987) para mapear derivações TAG superficiais de uma língua para outra. O mapeamento é realizado por meio de um léxico bilíngüe que associa diretamente árvores fonte e alvo por meio de ligações entre itens lexicais e seus argumentos. De modo simplificado, cada entrada nesse léxico bilíngüe contém regras para o mapeamento entre a sentença na LF e a sentença na LA.

10.1.4. TA baseada em restrições

O paradigma baseado em restrições (*Constraint-Based Machine Translation*, ou CBMT) permite definir restrições em vários níveis de descrição lingüística, por exemplo, para os itens lexicais. Entre as abordagens que utilizam esse paradigma, estão os sistemas de TA que combinam a LFG (Kaplan e Bresnan, 1982), com restrições sobre os itens lexicais, como o LFG-MT (Kaplan et al., 1989)[24]. Nesse sistema, as operações de mapeamento requeridas na transferência são executadas por equações de transferência baseadas em restrições que relacionam estruturas-f (estruturas funcionais da LFG) fonte e alvo.

10.1.5. TA baseada em princípios

Sistemas PBMT (*Principle-Based Machine Translation*) são uma alternativa aos sistemas RBMT, nos quais as regras são substituídas por um pequeno conjunto de princípios que envolvem fenômenos morfológicos, gramaticais e lexicais, de um modo geral. Um exemplo de construção derivada de princípios gerais é a construção da voz passiva, conforme descrito por Berwick (1991)[25]. Como não existe uma única regra de mapeamento entre duas línguas naturais para a voz passiva, é comum utilizar-se um conjunto de princípios que definem as operações morfológicas e sintáticas necessárias.

O *Princitran* é um sistema PBMT (Dorr et al., 1995)[26], baseado nos princípios sintáticos da Teoria da Regência e Ligação (*Government-Binding,* ou GB – Chomsky, 1981) e nos princípios semântico-lexicais da LCS (Dorr, 1993). Nesse sistema, a construção de estruturas é adiada até que as descrições satisfaçam os princípios lingüísticos.

O paradigma PBMT é complementar às abordagens KBMT e EBMT, no sentido de que ele provê uma cobertura ampla para muitos fenômenos lingüísticos, mas lhe falta conhecimento mais profundo sobre o domínio de tradução.

---

[23] Abeillé, A.; Schabes, Y.; Joshi, A.K. (1990). Using Lexicalized Tags for Machine Translation. In *Proceedings of Thirteenth International Conference on Computational Linguistics (COLING – 90),* pp. 1-6. Helsinki, Finland. Apud (Dorr et al., 2000), p. 23.

[24] Kaplan, R.; Netter, K.; Wedeking, A.Z.J. (1989). Translation by Structural Correspondence. In *Proceedings of Thirteenth International Conference on Computational Linguistics (COLING – 90).* Helsinki, Finland. Apud (Dorr et al., 2000), p. 19.

[25] Berwick, R. C. (1991). Principles of Principle-Based Parsing. In R.C. Berwick, S.P. Abney, and C. Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics,* pp. 1-37. Kluwer Academic Publishers. Apud (Dorr et al., 2000), p. 26.

[26] Dorr, B.J.; Lin, D.; Lee, J.; Suh, S. (1995). Efficient Parsing for Korean and English: A Parameterized Message Passing Approach. *Computational Linguistics*, 21(2), pp. 255-236. Apud (Dorr et al., 2000), p. 27.

10.1.6. TA *shake and bake*

O S&BMT (*Shake & Bake Machine Translation*) (Beaven, 1992)[27] é um dos paradigmas de tradução mais recentes. Ele utiliza regras de transferência como mecanismo para realizar a tradução, mas enquanto o mapeamento entre itens lexicais é realizado por meio de regras de transferência padrão, o algoritmo para combinar esses itens para formar uma sentença na LA não é convencional (Dorr et al., 2000).

As regras de transferência são definidas com base em entradas lexicais bilíngües, que relacionam itens monolingües. Após a análise da sentença da LF, suas palavras são mapeadas em palavras da LA por meio das entradas bilíngües. O algoritmo que combina as palavras na LA tenta ordená-las baseando-se nas restrições sintáticas da LA.

Para construções complexas, como os casos de troca de núcleo, diferentemente da abordagem por transferência simples, o paradigma S&BMT é capaz de construir regras de mapeamento não composicionais selecionando as palavras na LA a partir de um léxico bilíngüe e tentando diferentes ordenações para essas palavras (*shake*) que satisfaçam todas as restrições sintáticas, até que a sentença seja produzida (*bake*).

Essas regras formam a base para a transferência entre as entradas lexicais na LF e LA. A idéia central desse paradigma é que, uma vez que os elementos bilíngües identifiquem corretamente os índices das entradas lexicais, um algoritmo S&BMT pode combiná-los. O principal benefício dessa abordagem é que os léxicos bilíngües precisam somente especificar o conhecimento contrastivo entre duas línguas; as gramáticas monolingües usadas para o *parser* e geração se responsabilizam pelo restante (Dorr et al., 2000). A desvantagem dessa abordagem é que a geração é um problema NP-completo, ou seja, não há um algoritmo eficiente para geração de uma estrutura S&BMT.

10.2. Paradigmas empíricos

Os paradigmas empíricos são os que utilizam pouca ou nenhuma teoria lingüística no processo de tradução. Em geral, eles indicam técnicas experimentais para especificar o mecanismo de tradução apropriado ao contexto em foco. Esses paradigmas passaram a ser bastante explorados nos últimos anos devido ao grande avanço de hardware e à disponibilidade crescente de recursos eletrônicos significativos (dicionários, corpora de textos bilíngües e monolíngües, etc.), componentes essenciais para o sucesso da investigação empírica.

10.2.1. TA baseada em estatística

Sistemas baseados em estatística (*Statistical-Based Machine Translation*, ou SBMT) utilizam técnicas estatísticas ou probabilísticas que contemplam as tarefas lingüístico-computacionais em foco na tradução (por exemplo, a desambigüização lexical).

A idéia dessa abordagem é que a tradução seja realizada por meio de dados estatísticos extraídos automaticamente de corpora de textos bilíngües paralelos. Alguns exemplos de dados que podem ser obtidos a partir da análise desses corpora são:

- probabilidade de uma sentença fonte ocorrer no texto-alvo;

[27] Beaven, J. Shake and Bake Machine Translation. In *Proceedings of Fourteenth International Conference on Computational Linguistics*. Nantes, France, pp. 603-609. Apud (Dorr et al., 2000), p. 28.

- probabilidade de uma palavra fonte ser traduzida como uma, duas ou mais palavras alvo;
- probabilidade de tradução de cada palavra em outra palavra da língua alvo;
- probabilidade da posição de cada palavra na sentença na língua fonte, quando essa posição não é a mesma que a da palavra na sentença alvo.

As probabilidades obtidas são utilizadas para calcular como uma sentença fonte pode ser traduzida em uma sentença alvo. Há diversas formas de realizar esse cálculo, por exemplo, ele pode ser baseado numa variante da Regra de Bayes, que equaciona o problema da tradução como a produção de uma saída que maximize um valor funcional (Pr(A|F)), este representando a importância de se manter a fidelidade ao texto original e a fluência do texto traduzido. Nesse método, a probabilidade de uma sentença alvo (A) ser a tradução de uma dada sentença fonte (F) é proporcional ao produto da probabilidade de que a sentença A seja uma construção legal na LA (fluência) e da probabilidade de que uma sentença na LF seja a tradução da sentença na LA (fidelidade). A seguinte equação expressa essa relação (Dorr et al., 2000):

$$Pr(A|F) = Pr(A) * Pr(F|A)$$

Um exemplo de sistema que utiliza esse modelo de processamento estatístico é o *Candide* (Brown, 1990), de tradução do francês para o inglês. Esse sistema considera que a probabilidade de qualquer palavra na sentença alvo ser parte de uma sentença legal depende das probabilidades de ocorrência das duas palavras anteriores e que a probabilidade de que a sentença inteira seja uma sentença legal é o produto de ocorrência de todas as triplas de palavras em um corpus de textos em inglês. Já a probabilidade de que uma palavra na LF seja uma tradução de uma dada palavra na LA depende somente da probabilidade da ocorrência da palavra em uma sentença alvo, de acordo com as probabilidades de alinhamento dos pares de sentenças no corpus.

Um dos problemas da abordagem estatística é a necessidade de corpora de textos substanciais e de boa qualidade, o que torna as traduções muito dependentes do domínio do corpus. Outro problema é que a única forma de melhorar a qualidade da tradução é melhorar a exatidão dos modelos probabilísticos da língua alvo e do processo de tradução, o que exige a adição de muitos parâmetros, além dos já requeridos pelos vários modelos disponíveis. Uma alternativa para amenizar ambos os problemas são os sistemas híbridos.

Uma descrição mais detalhada sobre o processo de criação de sistemas estatísticos de TA pode ser consultada em Knight (1999); em Borthwick (1997) são apresentadas três diferentes formas de modelar a TA estatística: *N-grams*, Árvore de Decisão e Entropia Máxima.

10.2.2. TA baseada em exemplos

Na abordagem EBMT (*Example-Based Machine Translation*), também chamada de **TA baseada em casos**, em vez de regras de mapeamento entre as línguas, utiliza-se um procedimento que tenta combinar o texto a ser traduzido com exemplos de traduções armazenados. A tradução é, portanto, por analogia com exemplos coletados a partir de traduções já realizadas, os quais são anotados com suas descrições superficiais, em um corpus bilíngüe alinhado.

Basicamente, a idéia é utilizar um algoritmo de unificação para encontrar o exemplo mais próximo da sentença de entrada, a partir do corpus bilíngüe. Esse procedimento resulta

num *template* de tradução, o qual pode, então, ser preenchido palavra-por-palavra, de acordo com as palavras da sentença de entrada.

A proximidade de cada exemplo com a sentença de entrada é determinada pela distância semântica entre as suas palavras, a qual pode ser calculada com base na distância entre essas palavras em uma hierarquia de termos e conceitos provida, em geral, por um *thesaurus* ou uma ontologia.

A combinação de frases requer pelo menos uma análise sintática básica das traduções paralelas, além de alguma análise semântica para determinar a proximidade da combinação. Assim, a tradução de sentenças exige também que a estrutura sintática da sentença fonte seja combinada com sentenças no corpus. A maioria dos sistemas EBMT não considera a combinação da sentença inteira, mas sim de algumas de suas partes, como sintagmas nominais ou preposicionais.

A exatidão e a qualidade da tradução dos sistemas que utilizam o paradigma EBMT dependem da existência de um bom conjunto de dados. A grande cobertura de divergências sintáticas e semânticas requerida pode resultar em um conjunto de informações cujo tamanho dificulta o armazenamento e as buscas. A combinação do paradigma EBMT com abordagens lingüísticas (especialmente com sistemas RBMT) permite diminuir o tamanho desse conjunto de informações. Alguns métodos para adicionar conhecimento lingüístico a sistemas EBMT são descritos por Brown (1999).

Uma das vantagens desse paradigma é que a qualidade da tradução pode melhorar de forma incremental à medida que os exemplos tornam-se mais completos, sem a necessidade de atualizar ou melhorar descrições lexicais ou gramaticais. Algumas complicações nesse modelo ocorrem, por exemplo, quando se tem um número diferente de exemplos e cada um combina com uma parte da sentença, mas as partes que eles combinam se sobrepõem (Arnold et al., 1993). Um exemplo de sistema que utiliza o paradigma EBMT é o *Pangloss* (Brown, 1996).

10.2.3. TA baseada em diálogo

Sistemas de TA baseados em diálogo (*Dialogue-Based Machine Translation*, ou DBMT) são voltados para usuários que são os autores do texto a ser traduzido. Esse tipo de sistema provê um mecanismo que estabelece um diálogo sobre a tradução com o usuário, permitindo que este desambigüise o texto de entrada e incorpore detalhes estilísticos para obter uma tradução de melhor qualidade. Sistemas DBMT são similares aos EBMT, no sentido de que uma representação básica do texto de entrada do usuário é construída e, à medida que ela é revisada por meio de diálogos iterativos com o usuário, são feitas tentativas para atualiza-la a partir de informações armazenadas em um banco de dados de traduções. Além da interação com o usuário durante o processo de tradução, como um mecanismo de desambigüização *on-line* guiado pelo usuário, essa interação pode ocorrer antes do texto de entrada ser repassado ao sistema, como uma forma de revisão prévia guiada pelo usuário.

Em alguns sistemas são codificadas informações de contexto, de modo que o sistema possa determinar a provável intenção do usuário. Usando essas informações, o usuário pode ser guiado por uma série de pontos de escolha, os quais permitem a construção de uma representação que é oferecida ao sistema como candidata à tradução.

Assim como os sistemas KBMT e EBMT, sistemas DBMT são mais voltados para domínios bastante restritos. Para domínios mais abrangentes, a quantidade de informações requerida é muito grande, o que dificulta o armazenamento e a busca.

Um exemplo de sistema desenvolvido com base no paradigma DBMT é o *ENtran* (Johnson e Whitelock, 1987), projetado para prover a construção de um texto de entrada

restrito que, traduzido, deixa vários fenômenos lingüísticos para serem processados pelo usuário.

10.2.4. TA baseada em redes neurais

A incorporação da tecnologia de redes neurais e abordagens conexionistas na TA (*Neural-Based Machine Translation*, ou NBMT) é uma área de pesquisa relativamente nova. Essa tecnologia tem sido utilizada basicamente nas funções de *parser*, desambigüização lexical e aprendizado de regras de gramática, considerando-se subconjuntos bastante restritos das línguas. A manipulação de grandes vocabulários e gramáticas aumenta demasiadamente o tamanho das redes neurais e dos conjuntos de treinamento e, conseqüentemente, do tempo de treinamento.

Segundo Dorr et al. (2000), apesar das várias pesquisas sobre esse paradigma, nenhum sistema real de TA foi construído baseado somente na tecnologia de redes neurais, por isso, essa é considerada mais uma técnica auxiliar para a TA.

10.3. Paradigmas híbridos

Muitos paradigmas, principalmente os empíricos, apresentam dificuldades para manipular alguns aspectos do processo de TA. Por exemplo, sistemas estatísticos (SBMT) não manipulam dependências conceituais de longa distância, enquanto que sistemas baseados em exemplos (EBMT) dificilmente tratam estruturas sentenciais complexas. Assim, é reconhecida a necessidade de combinar paradigmas de forma a explorar as vantagens de cada um.

Um exemplo de abordagem híbrida comum consiste em se utilizar paradigmas lingüísticos para a análise automática de textos-fonte e paradigmas estatísticos ou baseados em exemplos para resolver as traduções frasais e as interdependências de constituintes.

Paradigmas estatísticos e probabilísticos devem predominar se o objetivo for obter robustez e grande cobertura de dados. Já se o objetivo for tratar de detalhadas nuances da língua, devem predominar os paradigmas fundamentais. A forma exata de como combinar os diferentes módulos em um sistema, no entanto, permanece uma questão em aberto.

Brown e Frederking (1995), por exemplo, propõem o uso de informações estatísticas para melhorar os resultados da TA fundamental, já Och e Weber (1998) propõem o uso de categorias e regras para melhorar a TA estatística, enquanto Al-Onaizan et al. (1999) utilizam a análise de dados estatísticos para adquirir conhecimento lingüístico de forma automática a partir de corpora bilíngües.

Exemplos de abordagens híbridas são: o sistema *Pangloss* (Brown, 1996), que utiliza os paradigmas EBMT e KBMT juntamente com o método por interlíngua; o sistema *Lingstat* (Barnett et al., 1994)[28], que utiliza o método de transferência para a tradução do japonês para o inglês, com o uso de uma gramática livre de contexto probabilística; e o sistema *Vermobil* (Vogel et al., 2000; Vogel et al., 2000b), também de TA por transferência, que apresenta um módulo auxiliar estatístico.

É importante ressaltar que um modelo híbrido pode envolver não somente a combinação de paradigmas de TA, mas também a combinação de métodos de TA (sistemas *multi-engine*).

[28] Barnnet, F.; Cant, J.; Demedts, T.D.; Gates, B.; Hays, E.; Ito, Y.; Yamron, J. (1994). LINGSTAT: State of the System. In *ARPA Workshop on Machine Translation*. Vienna, Virginia. Apud (Dorr et al., 2000), p. 32.

11. Conclusões e comentários finais sobre TA

O interesse pela área de TA existe desde a década de 40, porém, tem se intensificado nos últimos anos, em função da grande quantidade de informações disponíveis principalmente em meio eletrônico e da crescente necessidade de comunicação entre pessoas de diferentes línguas.

As pretensões iniciais para esses sistemas, bastante ambiciosas, foram adequadas de acordo com as limitações de hardware/software e, principalmente, as limitações de cunho lingüístico. Hoje, as maiores restrições aos sistemas de TA são impostas praticamente pela falta de solução computacional para diversos problemas lingüísticos.

Os diversos sistemas de TA existentes apresentam diferentes finalidades, domínios e abrangência, variando desde sistemas simples, de tradução de palavras individuais, a sistemas mais complexos, que consideram informações semânticas e contextuais. Geralmente sistemas simples são de finalidade mais geral, em domínios mais abertos, abrangendo um grande número de dados. Porém, apresentam pior desempenho do que sistemas mais complexos, os quais, devido à quantidade de informações de que necessitam, costumam ser limitados a domínios específicos, com objetivos bem definidos e menor abrangência. Essas limitações permitem que esses últimos sistemas obtenham resultados consideravelmente mais satisfatórios que os primeiros, em termos de qualidade das traduções.

Nos últimos anos, foi possível perceber, principalmente em sistemas de grande porte, uma forte preocupação com a consideração de informações sobre o significado dos textos a serem traduzidos, sejam elas de natureza semântica, pragmática, ou de senso comum, de modo a melhorar a qualidade das traduções. Nesse contexto, o método de TA por interlíngua é o que se mostra mais adequado, por permitir a representação do conhecimento de forma abstrata, independente de língua e possibilitar o desenvolvimento de ambientes multilíngües com uma complexidade relativamente baixa, se comparada à de outros métodos.

Pôde-se observar também, no desenvolvimento de sistemas de TA, uma tendência a considerar combinações de diferentes métodos e/ou paradigmas, resultando em abordagens híbridas, para obter traduções de melhor qualidade, como por exemplo, o uso do paradigma estatístico (SBMT) como processamento complementar ao realizado por métodos de transferência.

De um modo geral, apesar da considerável evolução da TA, seus resultados ainda precisam ser bastante aprimorados. Nesse sentido, mantém-se a dependência de teorias lingüísticas bem definidas, para diferentes línguas, e de estudos na Lingüística Computacional para descobrir meios de implementa-las.

Referências da Parte II

Al-Onaizan, Y. et al.. (1999). *Statistical Machine Translation*. Final Report. In *Johns Hopkins University 1999 Summer Workshop on Language Engineering*, Center for Speech and Language Processing, Baltimore.

Alfaro, C. (1998). *Descobrindo, Compreendendo e Analisando a Tradução Automática*. Monografia de Conclusão de Especialização. PUC, Rio de Janeiro.

Arnold, D.J.; Balkan, L.; Humphreys, R.L.; Meijer, S.; Sadler, L. (1993). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London.

Boitet, C. (1994). (Human-Aided) Machine Translation: A Better Future?, Grenoble.

Borthwick, A. (1997). *Survey Paper on Statistical Language Modeling*. Tech. Report, New York University, New York.

Brown, P.F. (1990). A Statistical Approach to Machine Translation. In *Computational Linguistics*, 16(2).

Brown, R.D. (1996). Example-Based Machine Translation in the Pangloss System. In *Proceedings of the 16th International Conference on Computational Linguistics* - COLING-96, pp. 169-174. Copenhagen, Denmark, August 5-9.

Brown, R.D. (1999). Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation* - TMI-99, pp. 22-32. Chester, UK, August.

Brown, R.D.; Frederking, R. (1995). Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* - TMI'95, pp. 221-239. Leuven, Belgium, July 5-7.

Carlson, L.; Nirenburg, S. (1992). World Modeling for NLP. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*. Trento, Italy.

Chomsky, N.A. (1957). *Syntatic structures*. Mouton, Hague.

Chomsky, N.A. (1981). *Lectures on Government and Binding*. Dordrecht, Foris.

Dorr, B.J. (1992). The use of Lexical Semantics in Interlingual Translation. In *Journal of Machine Translation*, 7:3, pp. 135-193.

Dorr, B. J. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge.

Dorr, B.J.; Jordan, P.W.; Benoit, J.W. (2000). A Survey of Current Paradigms in Machine Translation. In M. Zelkowitz (ed), *Advances in Computers*, Vol 49, pp. 1-68. Academic Press, London.

Farwell, D.; Wilks, Y. (1991). Ultra: A Mulit-lingual Machine Translator. In *Proceedings of MT Summit III*, Washington.

Fillmore, C. (1968). The case for case. In Bach, E. and Harms, R.T. (orgs.), *Universals in linguistic theory*, pp. 1-88. Rinehard and Winston, New York.

Hutchins, J. (1998). Translation Technology and the Translator. In *Machine Translation Review*. Norfolk.

Isabelle, P. (1987). Machine Translation at the TAUM Group. In *Machine Translation: The State of the Art*, pp. 247-318. Edinburgh University Press, Edinburgh.

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge.

Johnson, R.L.; Whitelock, P. (1987). Machine Translation as an Expert Task. In S. Nirenburg, ed., *Machine translation – Theoretical and methodological issues*, pp. 136-144. Cambridge University Press, Cambridge.

Joshi, A. K. (1987). Introduction to Tree Adjoining Grammar. In A. Manaster Ramer (ed.), *The Mathematics of Language*, pp. 87-114. J. Benjamins.

Kaplan, R.M.; Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations.* The MIT Press, Cambridge.

Kay, M. (1994). Machine Translation: The Disappointing Past and Present. In *Survey of the State of the Art in Human Language Technology*. Xerox Palo Alto Research Group, California.

Knight, K. (1999). *A Statistical MT Tutorial Workbook. In Johns Hopkins University 1999 Summer Workshop on Language Engineering*, Center for Speech and Language Processing, Baltimore.

Mateus, M.H.M. (1995). Tradução automática: um pouco de história. In M. H. M. Mateus e A. H. Branco (orgs.), *Engenharia da Linguagem*, pp. 115-120. Edições Colibri, Lisboa.

Mitamura, T.; Nyberg, E.H.; Carbonell, J.G. (1991). An Efficient Interlingua Translation System for Multi-lingual Document Production. In *Proceedings of Machine Translation Summit III*, Washington D.C, July 2-4.

Nirenburg, S. (1987). Knowledge and choices in machine translation. In *Machine translation – Theoretical and methodological issues*, pp. 1-15. Cambridge University Press, Cambridge.

Nirenburg, S.; Raskin, V.; Tucker, A.B. (1987). The structure of interlingua in TRANSLATOR. In *Machine translation – Theoretical and methodological issues*, pp. 90-113. Cambridge University Press, Cambridge.

Nirenburg, S.; Goodman, K. (1991). *The KBMT Project: A case study in Knowledge-Based Machine Translation*. Morgan Kaufmann Publishers, California.

Nyberg, E.H.; Mitamura, T. (1992). The Kant System: Fast, Accurate, High-quality Translation in Practical Domains. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes.

Och, F.J.; Weber, H. (1998). Improving Statistical Natural Language Translation with Categories and Rules. *In COLING '98*.

Rosseta, M.T. (1994). *Compositional Translation*. Kluwer Academic Publishers. Dordrecht, The Netherlands.

Santos, D. (1998). Um olhar computacional sobre a tradução. In *Revista Internacional de Língua Portuguesa*.

Schank, R. (1975). *Conceptual Information Processing*. North-Holland Publishing Company.

Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status, and Future Prospects. In J. Slocum (org.), *Machine Translation Systems*, pp.1-41. Cambridge University Press, Cambridge.

Uchida, H.; Zhu, M.; Senta, T.D. (1999)*. *The UNL, a Gift for a Millennium*. UNU/IAS/UNL Center. Tokyo.

UNL (2001)*. *The Universal Networking Language (UNL) Specifications*. UNU/IAS/UNL Center. Tokyo.

Uschold, M.; Gruninger, M. (1996). Ontologies: principles, methods and applications. In *The Knowledge Engineering Review*, Vol. 11:2, pp. 93-136.

Vogel, S.; Och, F.J.; Ney, H. (2000). The Statistical Translation Module in the Vermobil System. In *KOVENS*.

Vogel, S.; Och, F.J.; Tillmann, C.; Nieben, S.; Sawaf, H.; Ney, H. (2000b). Statistical Methods for Machine Translation. *In Verbmobil: Foundations of Speech-to-Speech Translation* pp. 377-393. Wolfgang Wahlster (ed.). Springer Verlag, Berlin.

Wahlster, W. (1993). Verbmobil, translation of face-to-face dialogs. In *Proceedings of the Fourth Machine Translation Summit*, pp. 127-135. Kobe.

Ward, N. (1999). Machine Translation. Chapter 20 of *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* by Jurafsky, D. and Martin, J.H., Prentice-Hall. Wong, S K.

* Disponíveis em http://www.unl.ias.unu.edu/

PARTE III – A Sumarização Automática de Textos: Principais Características e Metodologias

12. A Sumarização de Textos

A sumarização de textos, de um modo geral, é uma atividade comum na vida de qualquer pessoa de nível de escolaridade médio ou superior. Textos são, se não um objeto principal de trabalho, um instrumento auxiliar para atualização ou comunicação em qualquer esfera profissional ou social. Os sumários de textos, aqui tomados em sua acepção de *resumos*, são, por sua vez, também textos. Por essa razão, constituem igualmente objetos de comunicação. Com o aumento desmedido dos meios de comunicação e distribuição, observamos hoje um acúmulo excessivo de textos de diversas naturezas, o qual nos leva, via de regra, à incapacidade de consumi-los em sua íntegra. Em geral, a razão dessa incapacidade está em nossa falta de tempo, dada a diversidade de tarefas que abarcamos na sociedade moderna. Não é à toa que, freqüentemente, recorremos às principais manchetes dos jornais diários ou às sinopses das principais notícias, elaboradas por autores de nossa preferência. Freqüentemente também assinamos semanários, como a Veja, Isto É ou, ainda, revistas mais especializadas, como a Exame ou *The Economist*, as quais nos trazem temas atuais, porém resumidos, dos principais acontecimentos ou avanços econômicos.

Considerando o meio acadêmico, vemos novos campos de uso intenso de sumários: vestibulandos brasileiros, hoje, não se dão ao trabalho de recorrer a obras completas da literatura recomendada. Recorrem, ao contrário, aos resumos tão difundidos na última década, supondo que absorverão os principais aspectos da obra correspondente sem o "pênalti" da leitura da obra completa, permitindo-lhes o sucesso no exame visado. Cientistas comumente selecionam o vasto material de atualização científica primeiramente perscrutando seu título e seu sumário, pois estes são itens obrigatórios de complementação da divulgação científica.

Com o crescente uso da Internet, essa situação só foi evidenciada: *viajar* pelas páginas de notícias a fim de apreender *o que é essencial* exige tempo, capacidade de *identificar o que é relevante*, no grande volume de informações disponível, e capacidade de mentalizar, de forma coerente, o conteúdo essencial. Todos esses fatores remetem à capacidade de *sumarizar*.

Passamos a ver, portanto, a grande difusão da tarefa de sumarização, cujos objetivos podem ser classificados de duas formas: do ponto de vista do leitor – e, portanto, do usuário de um sumário – e do ponto de vista do produtor – e, portanto, de seu escritor. Este último cenário nos dá o foco da sumarização textual, como a tarefa de *escrita de um texto condensado*, o sumário, com o objetivo de *transmitir ou comunicar somente o que é importante* de uma fonte textual de informação. Como veremos adiante, essa perspectiva é importante, pois ela direcionará a modelagem automática, para a especificação dos principais processos de tomada de decisão.

As principais premissas da sumarização podem, assim, ser enumeradas como segue:

- Está disponível um texto, aqui denominado *texto-fonte*, que deve ser condensado.
- A afirmação de que o objeto a ser sumarizado constitui um texto implica, adicionalmente, a existência de
    a) uma idéia central – o tópico principal do texto – sobre a qual se constrói a trama textual (no ensino fundamental, aprendemos que o texto deve ser desenvolvido a partir de uma idéia – nossa idéia central);
    b) um conjunto de unidades de informação que, reconhecidamente, têm relação com a idéia central em desenvolvimento;

c) um objetivo comunicativo central que, implícita ou explicitamente, direciona tanto a seleção das unidades de informação quanto a seleção da forma como a informação será estruturada, para estabelecer a idéia pretendida.
d) um enredo, tecido em função das escolhas antes citadas, visando transmitir a idéia central de forma coerente, a fim de atingir o objetivo comunicativo pretendido.

- Tomando por base essa relação de conceitos, a principal premissa da sumarização de textos pode ser, assim, expressa como a tarefa de *identificar o que é relevante* no texto e, então, *traçar o novo enredo*, a partir do conteúdo disponível, *preservando sua idéia central*, sem transgredir o significado original pretendido.

A não transgressão do original constitui a principal restrição da sumarização:

Em que medida e com que parâmetros ela se impõe na tarefa de sumarização?

Apesar de serem vagos os conceitos acima e, logo, insuficientemente formais para estabelecer um modelo de produção de sumários, como falantes graduados em nível médio ou superior conseguimos identificar, mesmo que intuitivamente, as premissas acima como determinantes de dois parâmetros essenciais de julgamento da relação entre esses objetos textuais: (a) a relação do sumário com um texto-fonte (sendo este mais estendido e detalhado do que aquele) e (b) a qualidade do sumário. Como leitores, conseguimos reconhecer, e bem, o grau de proximidade entre ambos, a ponto de qualificar sumários como bons ou ruins dependendo de sua fidelidade ao que é essencial em sua fonte textual.

Como escritores, dificilmente parecemos guiados por um vínculo estrito e explícito, previamente estabelecido, para garantirmos a proximidade de um sumário com seu texto-fonte, muito embora sejamos capazes de estabelecê-lo, garantindo que o leitor possa reconhecer um sumário como o veículo de comunicação sucinta do que antes era expresso mais detalhadamente. Freqüentemente violamos, por exemplo, o objetivo comunicativo de um texto, ao produzir vários sumários, com diversas conotações. Isso é particularmente evidente em nosso meio acadêmico: ao submetermos um artigo a uma revista, primeiramente temos por objetivo convencer o editor de que o artigo deve ser publicado; ao termos o trabalho aceito, freqüentemente alteramos o sumário, agora visando o leitor da comunidade mais abrangente, de interessados no assunto em si. Assim, cada objetivo é determinante da forma como o enredo do sumário é construído. No primeiro caso, a ênfase no convencimento do leitor imediato pode fazer com que o conteúdo do mesmo seja menos técnico e, portanto, mais independente do próprio teor técnico do artigo. No segundo caso, a ênfase passa a ser na divulgação científica de seu conteúdo e, assim, convencer o leitor de que ele é válido deixa de ser relevante, passando a ser prioritário motivá-lo para a leitura do artigo completo. Essa conotação pode se relacionar, ainda, ao objetivo de atrair a atenção do maior número de leitores, implicando uma mudança radical da conotação anterior.

Outro exemplo mais comum de variadas formas de sumarização, porém, todas elas remetendo a um mesmo texto-fonte, é o de manchetes de jornais, que fazem parte de nossa vida diária. O Texto 1 da Figura 1, por exemplo, extraído de um jornal[29], pode ter as manchetes M1-M4, as quais ressaltam informações diferentes, atribuindo diversos graus de relevância ao conteúdo originário do mesmo texto-fonte. Vale notar, também, que o foco das manchetes M1-M3 é explícito no Texto 1: M1 e M3 são derivadas da sentença [1] e M2, da sentença [3]. Porém, o foco de M4 está implícito no texto, tendo sido inferido do conjunto de sentenças [8]-[10]. A inferência, neste caso, pode ser uma tentativa de recuperação do objetivo do escritor do texto (justificar o gasto com pesquisas; ressaltar a importância da pesquisa genética, etc.).

[29] Corpus jornalístico do NILC (Pinheiro e Aluísio, 2003).

Texto 1: [1] Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. [2] A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. [3] Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. [4] O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. [5] Assim, aumentaria a população de insetos refratários ao vírus. [6] A dengue provoca náuseas e dores de cabeça, articulações e músculos. [7] O tipo mais grave da doença, o hemorrágico, pode matar. [8] Em 1995, foram registrados 120 mil casos da doença no Brasil. [9] Em abril, o Ministério da Saúde anunciou um programa de combate à doença, que vai durar quatro anos e custar cerca de R$ 5 bilhões. [10] Cerca de 1250 municípios brasileiros, aproximadamente um em cada quatro, registraram casos de dengue.

Figura 1. Texto ilustrativo

M1: A contribuição da pesquisa genética ao combate à dengue
M2: A criação de um mosquito resistente ao vírus da dengue
M3: O combate à transmissão da dengue com a ajuda de mosquitos alterados geneticamente
M4: A pesquisa genética pode ajudar a minimizar os custos de combate à dengue

Assim como manchetes podem ser uma forma de sumário de um texto, várias outras formas, além de textos condensados, podem ser reconhecidas como sumários, cada uma delas envolvendo pressuposições, conteúdos e características diversos, prevalecendo, contudo, sua correspondência com as respectivas fontes.

Sumários são, assim, entendidos (e usados), hoje, como objetos autônomos de comunicação. A sumarização humana, por sua vez, pode ser definida como "a tarefa de redução do tamanho de um texto-fonte, com preservação de seu conteúdo mais relevante".

Identificadas as principais características da tarefa humana de sumarização, resta saber como a Sumarização Automática de textos (doravante, referenciada por sua sigla, SA) pode incorporá-las, para simular a produção automática de sumários textuais e garantir que a correspondência entre os resultados automáticos e os textos-fonte é, de fato, consistente. Distinguiremos, aqui, duas áreas igualmente importantes, para o projeto e desenvolvimento (P&D) dos sistemas computacionais dessa natureza, i.e., nossos *sumarizadores automáticos*[30]: a de modelagem de procedimentos para escolha e estruturação dos sumários a serem gerados automaticamente e a de avaliação dos mesmos, visando à avaliação do desempenho computacional. Nesse contexto, um sumarizador automático pode ser definido como

Um sistema computacional cujo objetivo é produzir uma representação condensada do conteúdo mais importante de sua entrada, para consumo por usuários humanos

Para isso, ele deve ser capaz de identificar, em um texto ou em uma representação conceitual do mesmo, o que é relevante, estruturando as unidades informativas correspondentes de modo a assegurar que o sumário será coerente e consistente. Segundo Mani (2001), essa caracterização distingue a SA de outras áreas correlatas, dentre as quais destacamos:

- Recuperação de Documentos, que, para uma certa "chave de busca", visa produzir uma coleção de documentos relevantes, sem necessariamente condensá-los;

[30] 'Sumarizador automático' será o termo usado para expressar, simplesmente, os sistemas computacionais que têm por objetivo sumarizar textos em língua natural.

- Indexação, que visa identificar termos convenientes para a recuperação de informação;
- Extração de informação, que não necessariamente tem a condensação de informação como restrição fundamental;
- Mineração de textos, cuja principal função é identificar, nos mesmos, informações singulares, e não necessariamente informações principais, como é o caso da recuperação e preservação da idéia central, na SA.

Assim, a modelagem de um sumarizador automático terá como principais restrições as mesmas da tarefa humana: os sumários devem ser textos condensados de uma fonte (em nosso caso, um texto ou sua representação conceitual) e, como tais, devem ter um enredo claro e progressivo, desenvolvido em torno de uma idéia central, a qual deve coincidir com a idéia central da fonte. Essas restrições imporão critérios claros para a averiguação do desempenho dos sumarizadores automáticos. Entretanto, a automação da tarefa não se restringe à mimetização do processo de escrita que normalmente é familiar aos falantes de uma língua, devido à sua complexidade: simular a reescrita de um texto, como faz o escritor humano, é um grande problema, pois envolve processos igualmente complexos, de interpretação do texto e representação (obrigatória) de somente uma parcela dele – aquela relacionada à sua idéia central. Essa complexidade será evidenciada ao descrevermos a abordagem fundamental de SA. Além desta, há um grande interesse, atualmente, pela abordagem empírica, que, diferentemente de incorporar modelos vinculados aos de comportamento humano na tarefa de sumarização, baseia-se em modelos matemáticos ou estatísticos para produzir resultados análogos.

O P&D em função dessas abordagens é, portanto, distinto: na primeira, o sumarizador automático deve incorporar modelos lingüísticos e/ou discursivos de interpretação e reescrita textual; na segunda, ele se baseia em modelos exatos de manipulação do conteúdo textual. Devido a essa diversidade, o processamento resultante caracteriza-se da seguinte forma:

- Abordagem fundamental: alto grau de representação simbólica do conhecimento lingüístico e textual e raciocínio lógico baseado em técnicas simbólicas, para estruturação e reescrita do sumário;
- Abordagem empírica: processamento prioritariamente baseado em reconhecimento de padrões derivados de informações ou distribuições numéricas. São usadas técnicas empíricas e/ou estatísticas para a extração dos segmentos textuais relevantes.

Em ambos os casos, consideram-se as principais premissas da sumarização textual, antes delineadas. Entretanto, é preciso trabalhar com as informações textuais, distinguindo-as e delimitando-as, para reconhecer tanto seu grau de relevância quanto seu inter-relacionamento, características que irão subsidiar as escolhas automáticas. Em geral, as informações textuais são associadas a unidades de conteúdo (ou unidades informativas), identificadas como unidades mínimas de significado no texto. Essas unidades podem expressar diversos níveis de detalhe e remeter a diversos contextos textuais (por sua localização), que sugiram diferentes mecanismos de compreensão e apreensão da mensagem contida no texto. Assim, a delimitação de unidades informativas simples contribui para a delimitação de contextos variados, tornando imprescindível distinguirem-se os limites textuais. É comum associar-se a uma sentença simples, por exemplo, um único significado, ou a um parágrafo, um único tópico do discurso. No Texto 1, por exemplo, distinguimos segmentos relacionados a diferentes tópicos, como demonstram as manchetes ilustrativas. Temos, nesse caso, três tópicos distintos, sendo o terceiro construído pela composição das sentenças [8]-[10].

As estruturas textuais irão contribuir igualmente para estabelecer o enredo textual e, assim, identificar as unidades que devem compor o sumário. Por isso, é importante definir a

granularidade das unidades informativas, para sua delimitação. Há diferentes abordagens nesse sentido, como veremos a seguir.

13. A Sumarização Automática de Textos

As possibilidades de Sumarização Automática delineadas na seção anterior indicam o grande problema da Sumarização Automática: produzir sumários que, mesmo diversos de seus textos-fonte, reflitam sua interdependência. Duas características são essenciais nesse contexto: (a) sumários remetem, necessariamente, a seus textos-fonte; (b) sumários devem ser construídos de modo a não haver perda considerável do significado original, apesar de conterem menos informações e poderem apresentar diferentes estruturas, em relação a suas fontes. Assim, sumários são textos produzidos a partir de textos ou de suas correspondentes representações, podendo servir de indexadores ou substitutos dos mesmos. Essa distinção leva à sua classificação como sumários *indicativos* ou *informativos*, respectivamente (Mani e Maybury, 1999)[31]. Sumários indicativos não podem substituir os textos-fonte, pois não necessariamente preservam o que aqueles têm de mais importante, em termos de conteúdo e estrutura, transmitindo somente uma vaga idéia daqueles. Sumários informativos, ao contrário, contêm todos os seus aspectos principais, dispensando, por isso, sua leitura (são chamados autocontidos, nesse caso).

Em função dessa classificação, distinguem-se também sua funcionalidade e a forma de avaliar sua qualidade: sumários indicativos podem ser utilizados na classificação de documentos bibliográficos, de um modo geral, indicando seu conteúdo e agilizando o acesso às informações relevantes. Nesse caso, eles servem de *indexadores*. Sumários informativos, por serem autocontidos, servem de meios de informação, porém, apresentam uma relação mais complicada com seu texto-fonte, pois de seu objetivo dependerá bastante a avaliação sobre o quanto ele atende às necessidades do usuário. A utilidade dos primeiros sumários é mais clara e sua função mais limitada do que a dos últimos. Essas características permitem uma avaliação mais robusta de sua funcionalidade e qualidade. De um modo geral, também é mais fácil produzir automaticamente sumários indicativos do que informativos. Entretanto, ambos podem servir a diversas aplicações, ressaltando-se, especialmente, a área de Recuperação de Informação, muito importante nos dias de hoje.

Além da diferença funcional, os sumários também são comumente classificados pelo modo como são obtidos. Sparck Jones (1993a) classifica-os como *extracts* ou *abstracts,* fazendo a correspondência com o que ela chama, respectivamente, de *extração textual* e *condensação de conteúdo* (Sparck Jones, 1997), sendo este o processo correspondente à reescrita textual, antes citada. Essas formas remetem às abordagens descritas como empíricas e fundamentais, respectivamente, também denominadas abordagens baseadas em corpus e abordagens baseadas em conhecimento profundo (*knowledge-rich approaches*). Ambas as abordagens podem, ainda, explorar a estruturação do discurso, porém, distinguem-se na forma como a estrutura do mesmo é manipulada. Elas ainda delineiam arquiteturas típicas, que serão descritas abaixo.

Nesse texto, usamos o termo geral 'sumário' para nos referir tanto a *extracts* quanto a *abstracts.* Quando o sumário for derivado da metodologia fundamental, especificamente, associaremos esse termo ao próprio termo em inglês, *abstract*; quando ele for derivado da metodologia empírica, usaremos, simplesmente, a tradução literal para *extract*, i.e., 'extrato'.

De um modo geral, a SA pode ser expressa por três processos (Mani e Maybury, 1999), como ilustra a Figura 2: análise, transformação e síntese. A análise consiste em extrair

[31] Nessa classificação inclui-se ainda um terceiro tipo – o de sumário crítico (que seria uma resenha do texto) – não considerado nesse capítulo.

uma representação computacional do texto-fonte; a transformação consiste em manipular essa representação a fim de produzir a representação do sumário. Finalmente, a síntese consiste em realizar linguisticamente esta última estrutura, produzindo o sumário, propriamente dito. Veremos que esses processos tomam diferentes formas, dependendo da abordagem considerada.

13.1. A SA Extrativa: Métodos Estatísticos e/ou Empíricos

A SA começou a ser explorada no final da década de 50, com a utilização expressiva de técnicas estatísticas de extração de conhecimento lingüístico dos textos-fonte. Luhn (1958), por exemplo, sugeriu o uso de informações estatísticas derivadas do cálculo da freqüência das palavras e de sua distribuição no texto para calcular uma *"medida relativa de significância"*[32]. Utilizou, para isso, tanto a granularidade individual (palavras), quanto a sentencial: sentenças mais significativas teriam *pesos maiores* e, assim, seriam escolhidas para compor o que ele chamou de *auto-abstract*. Adicionalmente, palavras mais significativas (e, portanto, de maior freqüência) correspondiam às atuais palavras-chave, tão conhecidas como representantes do conteúdo textual.

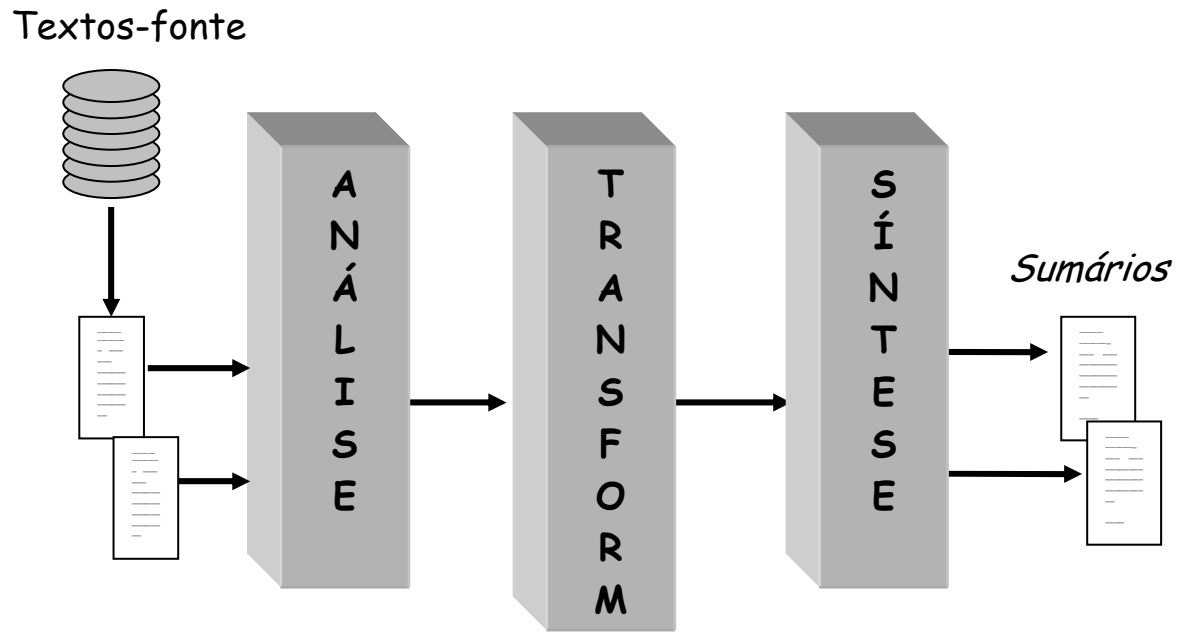


Figura 2. Arquitetura geral de um sumarizador automático

Dez anos depois, Edmundson (1969) propôs avanços sobre a metodologia de Luhn, denominando seu método de *"seleção computacional de sentenças com maior potencial de transmitir ao leitor a substância do documento"*. Além das distribuições sugeridas por Luhn, ele considerou também *"palavras pragmáticas"* (*cue words*), as quais hoje denominamos, simplesmente, *palavras sinalizadoras* (da importância do conteúdo textual, como *muito importante, significativa*, etc.), títulos e cabeçalhos, além de indicadores da localização do segmento na estrutura textual[33]. Seu sistema extrativo foi parametrizado de acordo com a

[32] Frases em itálico e entre aspas são traduções dos termos literais encontrados nas obras citadas.
[33] Sentenças mais importantes estariam, por exemplo, no começo e no fim de um documento ou seriam as primeiras e últimas de um parágrafo (Baxendale, 1958) ou, ainda, estariam logo abaixo de um titulo de seção.

influência de todos esses componentes, de acordo com uma combinação ponderada de características (normalmente, termo usado em inglês: *features*). Para essa ponderação, ele já fazia uso de dicionários eletrônicos, para reconhecer os segmentos textuais relevantes para compor os extratos[34]. Também significativa foi sua avaliação, comparando os extratos automáticos com manuais e sugerindo desempenho melhor do que aquele baseado, simplesmente, na distribuição de freqüência. Entretanto, o foco desse trabalho não estava, propriamente, na garantia de progressão textual, mas sim na reprodução automática da tarefa de indicação de seu conteúdo (*document screening*), mesmo que não coeso ou coerente.

Significativo foi também o trabalho de Pollock e Zamora (1975), que sugeriu a necessidade de se restringir domínios (ou assuntos) para melhorar os resultados de métodos extrativos de SA, propondo, em adição aos trabalhos anteriores, o cruzamento de sentenças com o título da obra, para determinar aquelas significativas para um extrato. Vale notar que, neste caso, era necessário haver um título associado ao texto-fonte, para implementar o método.

De um modo geral, essas foram as obras clássicas de SA que deram origem ao que temos, hoje, de mais moderno em SA extrativa. Por longo tempo, entretanto, a exploração de métodos nessa linha ficou estagnada, devido à impossibilidade técnica para implementá-los (limitações de *hardware* e *software*, mas também de disponibilidade de recursos eletrônicos, como dicionários ou repositórios lingüísticos de grande porte). Na década de 90, vemos o ressurgimento do interesse por essa abordagem: os computadores passaram a ser de uso geral, suas memórias baratearam e recursos lingüísticos expressivos, como etiquetadores morfossintáticos e *stemmers*, tornaram-se disponíveis para o processamento textual. O conhecimento sobre manipulações estatísticas mais elaboradas pôde, assim, ser explorado para a SA de textos de domínios e gêneros variados, dando origem à metodologia baseada em corpus e caracterizando mais propriamente as diversas formas de transformação de um dado de entrada, para a o produção dos extratos. Assim, a partir da arquitetura geral (Figura 2), caracterizou-se a abordagem empírica como um processo de manipulação numérica/estatística de informações, ilustrado na Figura 3.

[34] Aparentemente, nesse trabalho temos a primeira referência a *extratos* como sumários produzidos automaticamente pela metodologia de *extração de segmentos textuais*. O mesmo termo faz alusão às "*porções de um documento selecionadas para representar seu todo",* segundo Weil: Weil, B.H. (1970), Standards for writing abstracts. *Journal of the American Society for Information Science* 22(4): 351-357. (apud Pollock e Zamora, 1999, p. 43).

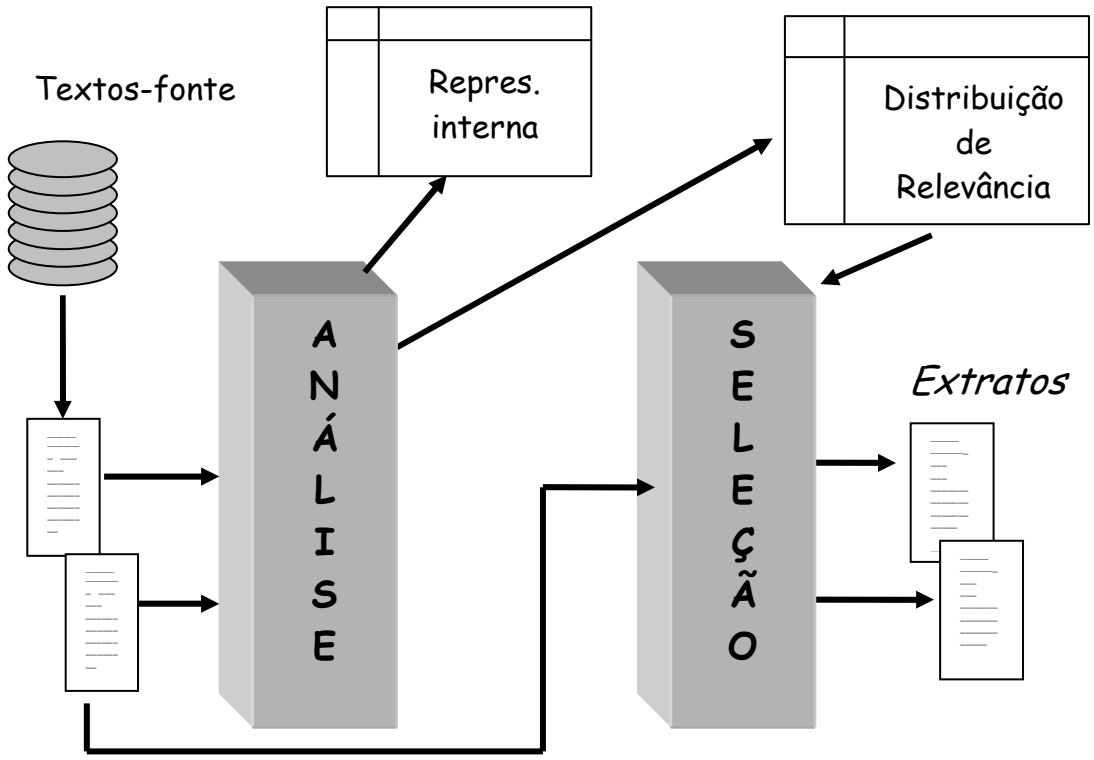


Figura 3. Sumarizador automático empírico

É na fase de análise textual que se exprime o avanço da SA extrativa: os procedimentos fundamentam-se nas distribuições clássicas de Luhn, Edmundson e Baxendale, porém, tornaram-se bastante sofisticados mais recentemente. A análise corresponde a algum tipo de esquadrinhamento e preparação do texto, para a manipulação das informações na fase de transformação. É comum termos, como sua representação interna, um vetor de características, sendo estas correspondentes aos aspectos significativos do texto-fonte. Por exemplo, podem ser consideradas somente palavras de significado, ou palavras de classe aberta (substantivos, adjetivos, verbos e advérbios), as quais podem dar origem aos radicais (*stems*) correspondentes, anotados com sua categoria sintática ou morfológica. São removidas dos textos-fonte, portanto, as palavras de classe fechada (pronomes, conjunções, artigos, preposições), assim como palavras cujo significado seja irrelevante para o contexto (em geral, estas são dependentes do domínio em foco). É criada, uma lista com essas palavras, chamada lista de *stopwords.* Esta etapa da análise da entrada do sumarizador é chamada, assim, remoção de *stopwords.* As informações restantes darão origem à distribuição de relevância, que indicará os segmentos textuais selecionados para compor o extrato. Hoje também é comum utilizar grandes volumes de dados textuais, para treinar o sumarizador a reconhecer informações significativas de textos em domínios específicos. As informações de entrada, neste caso, são classificadas segundo sua significância no contexto. Essa classificação serve de parametrização do sistema, para a escolha de segmentos durante a SA de textos-fonte de mesmo domínio e gênero que os utilizados no treino.

Os métodos empíricos adotam, em geral, a extração como processo fundamental: uma vez identificados como relevantes, segmentos inteiros de texto são extraídos do texto-fonte e integralmente incorporados ao extrato em construção, na mesma ordem em que eles se apresentam no texto-fonte. Devemos lembrar, portanto, que a síntese, neste caso, consiste simplesmente da justaposição dos segmentos considerados. Como é comum usar-se a medida sentencial de delimitação dos segmentos textuais (sob a suposição de que sentenças são a

unidade mínima de significado), a escolha das sentenças visa à garantia de preservação do enredo (e, logo, da coerência textual).

Tipicamente, os métodos de identificação de segmentos relevantes calculam a significância de cada sentença em um texto-fonte por seu peso e, então, selecionam aquelas com maior peso (acima de um limite mínimo) para compor o extrato, incorporando os parâmetros clássicos de identificação e seleção das mesmas (Luhn, 1958; Baxendale, 1958; Edmundson, 1969). Porém, eles manipulam grandes volumes de dados textuais para extrair as principais *features* para a identificação e seleção da sentenças, considerando domínios e gêneros particulares para melhorar o desempenho dos sumarizadores. As propostas mais recentes exploram, ainda, os papéis semânticos de cada unidade informativa (p.ex., Paice e Jones, 1993), as relações retóricas delineadas pelo inter-relacionamento de diversas unidades informativas (p.ex., Marcu, 1997a, 2000; Miike et al., 1994) e a similaridade estrutural entre sentenças (p.ex., Salton et al., 1997).

Um dos trabalhos mais importantes nessa linha é o de Kupiec et al. (1995), que propõe um sumarizador extrativo que, primeiramente, deve ser treinado para reconhecer as características textuais que parametrizam o sistema, i.e., aquelas que servem de base para a determinação da significância dos segmentos textuais, para inclusão no extrato. A hipótese principal dessa proposta é que a formulação de regras de ponderação da significância de sentenças de um texto é heurística e empírica por natureza. Assim, ela depende do treinamento do sistema sobre um corpus específico, o corpus de treino. O sistema é baseado em um classificador estatístico (bayesiano) (Mitchell, 1997) que agrupa as potenciais *features* a partir da comparação do conteúdo de textos-fonte com o conteúdo de seus respectivos extratos, construídos manualmente. Como resultado, é possível elencar um grupo de *features* ou de critérios de ponderação, para manipular as sentenças dos textos a sumarizar. A restrição, aqui, é que estes sejam similares àqueles do corpus de treino, devido à dependência de gênero e domínio, na fase de treinamento. Após vários experimentos, Kupiec et al. fixaram um conjunto de *features* consideradas significativas para seus dados (um corpus de textos sobre engenharia), que compreende o comprimento da sentença, a existência de sintagmas sinalizadores, a localização de sentenças no texto-fonte (inicio, meio ou fim de parágrafos), um conjunto de palavras "temáticas" (as mais freqüentes, neste trabalho, são consideradas temáticas) e a ocorrência de substantivos próprios. Embora essas *features* tenham sido derivadas de um corpus específico, não é difícil constatar que a maioria delas é comum a outros domínios e, logo, a proposta é bastante abrangente.

Uma vez construídas as classes de *features,* são calculadas as probabilidades de todas as sentenças de um texto-fonte, de serem incluídas em um extrato. Essas probabilidades irão indicar, portanto, sua seleção (ou exclusão) do texto final. Adicionalmente, a decisão de inclusão depende, também, da taxa de compressão desejada pelo usuário: essa taxa corresponde ao *volume de redução* do texto-fonte. Considerando-se que seu tamanho possa ser medido por número de sentenças, por exemplo, ela corresponde ao número de sentenças que deverão ser *excluídas* do sumário final. Assim, ela pode ser definida pela fórmula geral (TC = Taxa de Compressão):

$$TC = 1 - (\text{tamanho do extrato} / \text{tamanho do texto-fonte})$$

Atualmente, é comum se estabelecer a TC, principalmente na abordagem empírica, para evitar que sentenças muito longas resultem em textos pouco condensados ou, ao contrário, para estabelecer a diversidade pretendida pelo usuário, na produção de sumários. Em geral, na SA consideram-se sumários que correspondam a 10-20% dos textos-fonte e, portanto, que tenham uma taxa de compressão de 80 a 90% desses.

O trabalho de Kupiec et al. foi o marco responsável pelo *boom* da exploração de técnicas extrativas ainda mais robustas, estabelecendo a área hoje conhecida como SA baseada em corpus: métodos estatísticos de extração operam sobre sumarizadores treináveis a partir de corpora robustos de textos. A SA passou a ser, assim, um problema de classificação estatística: o objetivo é buscar uma função que calcule a probabilidade de uma sentença ser incluída no extrato (e, logo, que expresse sua significância), combinando características diversas. O problema, aqui, é determinar a contribuição relativa de diferentes *features,* condição altamente dependente do gênero textual (Mani e Maybury, 1999): textos científicos, por exemplo, podem concentrar informações relevantes no *abstract* e nas conclusões. Heurísticas distintas, derivadas da classificação das informações nos corpora, podem resolver esse problema. Essa possibilidade de treinar um sumarizador com base em textos-fonte específicos, para melhorar seu desempenho, trouxe também novas perspectivas para a avaliação dos resultados automáticos.

Teufel e Moens (1999) estendem o método de Kupiec et al., adicionando à classificação probabilística a função retórica de cada sentença, associada à estrutura do discurso. Seu trabalho diverge, assim, na forma de análise: ainda é preciso esquadrinhar o texto-fonte, para produzir uma distribuição de seus segmentos. Porém, também são identificados os papéis retóricos de cada sentença no texto. A extração da distribuição retórica de um texto-fonte é baseada em sua macro-estrutura: as categorias distintas de informação que caracterizam os segmentos mais genéricos do texto são responsáveis por indicar a funcionalidade de cada segmento. Por exemplo, para textos científicos, os macro-componentes podem incluir *problema, propósito, metodologia, resultados, conclusões, trabalho futuro*, etc. Também nesta fase Teufel e Moens usam o classificador bayesiano.

Seguindo essa abordagem híbrida, foram desenvolvidos métodos mais sofisticados, que incorporam ao tratamento numérico das informações textuais também a identificação e o processamento empírico de informações lingüísticas e discursivas dos textos-fonte. Esses métodos são diferenciados por sua abordagem baseada na estruturação do discurso (Mani e Maybury, 1999). A proposta de Barzilay e Elhadad (1997) é um exemplar disso: seu sumarizador explora a *coesão lexical* (i.e., o encadeamento de itens lexicais no texto), identificando nos textos-fonte as possíveis *cadeias lexicais*. Aquelas cadeias mais fortemente conectadas indicam as sentenças significativas para compor o extrato.

O trabalho fundamental dessa proposta de automação é puramente lingüístico, remetendo à coesão textual (Halliday e Hasan, 1976) e ao uso da repetição lexical para determinar os graus de coesão (Hoey, 1991). É baseada na proposta manual de cômputo das cadeias lexicais de Morris e Hirst (1991), havendo se tornado factível automaticamente pela disponibilidade de fontes robustas de conhecimento, tais como (a) um thesaurus que pudesse indicar os elos entre diversas palavras – a WordNet (Miller, 1995); (b) um etiquetador morfológico – que associa etiquetas a cada palavra, indicando sua categoria morfológica; (c) um *parser,* para identificar grupos nominais (envolvendo substantivos e adjetivos) e (d) um algoritmo de segmentação textual, responsável por delimitar, no texto-fonte, os segmentos que indicam as cadeias léxicas mais fortes.

Barzilay e Elhadad consideram somente substantivos e compostos nominais, para compor cadeias lexicais. O cômputo do relacionamento semântico das cadeias de palavras é feito de diversas formas: pela identificação de palavras idênticas ou palavras com mesmo significado; por sinonímia; por relações ontológicas de herança ou "parentesco". No primeiro caso, incluem-se a hiperonímia ou hiponímia (relações de superclasses ou subclasses, respectivamente). Por exemplo, *carro* e *Toyota* têm seus significados ontologicamente relacionados. No segundo caso, incluem-se relações de mesmo nível (também chamadas paratáticas). Por exemplo, a existente entre *caminhão* e *carro*.

A fase de análise dos textos-fonte compreende seu pré-processamento, pela seleção das palavras candidatas e segmentação do texto-fonte em tópicos, e a construção das cadeias lexicais, propriamente dita, que envolve a identificação das relações ontológicas, entre as palavras. Finalmente, a síntese para a produção dos extratos baseia-se em três heurísticas distintas, para identificar as sentenças que contêm as cadeias lexicais fortes: a que focaliza a primeira ocorrência das sentenças no texto-fonte, a que identifica as sentenças que possuem os membros mais representativos e a que se concentra na significância do tópico indicado pelas sentenças. Esta última heurística sugere que um leitor pode identificar melhor o tópico de um texto simplesmente identificando suas cadeias lexicias mais representativas.

O maior problema dessa abordagem é identificar as palavras polissêmicas da língua natural: não há repositório eletrônico capaz de definir as acepções mais prováveis para casos ambíguos, pois elas são dependentes do contexto, o qual é variável. Assim, várias cadeias lexicais podem ser derivadas de uma única construção, dificultando, e mesmo piorando, a tarefa de identificação das informações relevantes. Outro problema, extensivo às demais abordagens empíricas, é introduzido pela impossibilidade de resolver anáforas ou de controlar o nível de detalhe dos extratos resultantes, pois não é feito o tratamento interpretativo do material indicado pelas cadeias lexicais. Essa questão, via de regra, só poderá ser adequadamente tratada pela abordagem fundamental.

Propostas como as descritas evidenciam a grande variedade de abordagens extrativas, várias delas recorrendo a técnicas de aprendizado e treinamento automáticos com base em grandes corpora de textos, resultando em técnicas mais robustas, porque mais informadas, quando comparadas aos métodos extrativos mais simples. É importante notar que elas sugerem a manipulação numérica, em geral estatística, de componentes textuais, considerando medidas que, *implicitamente*, incorporam características lingüísticas e a experiência de sumarizadores humanos. De fato, na tarefa de identificação e cópia de material dos textos-fonte para produzir os extratos, as métricas da SA extrativa modelam, sobretudo, as adotadas por sumarizadores profissionais (p.ex., Borko e Bernier, 1975; Cremmins, 1996), e, logo, estão próximas à tarefa de *sumarização* profissional, área clássica, anterior ao uso do computador para simulá-la. A manipulação numérica inicial foi, assim, incrementada com a disponibilidade de recursos lingüísticos mais abrangentes e, conseqüentemente, de sumarizadores automáticos mais sofisticados e robustos. Por esse motivo, hoje a adoção de métodos empíricos é muito promissora, principalmente ante a urgência de se processar grandes volumes de informações textuais disponíveis eletronicamente. Assim é que podemos encontrar ferramentas de SA na Internet (como a do AltaVista) ou em ambientes de edição de textos (como o AutoResumo do MS Word™).

Os avanços da SA extrativa evidenciam ainda uma área inovadora e igualmente importante: a de avaliação. Em geral, usam-se "*gold standards*" (Kupiec et al., 1995) – padrões de referência definidos por especialistas humanos em sumarização textual – tanto para o treinamento dos sistemas quanto para sua avaliação.

Durante a fase de estagnação da SA extrativa, observamos a instalação da abordagem fundamental para a SA de textos, sobretudo a partir das idéias de Chomsky (1965): a modelagem computacional dos processos de compreensão e apreensão da estrutura textual, a fim de reescrever o texto-fonte de forma condensada, pôde ser formalizada a partir de gramáticas livres de contexto, responsáveis por analisar sintaticamente (*parsing*) os textos-fonte (de um domínio particular), para produzir sua representação conceitual.

13.2. A SA baseadas em conhecimento profundo: métodos fundamentais

Os principais problemas da abordagem fundamental (ou analítica) estão na forma como é identificada e sintetizada a informação relevante: a Figura 4 sugere a simulação do próprio processo humano de sumarização, composto da compreensão do enunciado do texto-fonte, com posteriores condensação de conteúdo e reescrita textual, conceitos já introduzidos no início desta seção, denominados por Hovy e Lin (1997) de *reescrita* e *fusão* de vários conceitos em um número menor de conceitos. Portanto, são três as etapas de SA (Sparck Jones, 1993b): a construção de uma representação do significado a partir do texto-fonte (repres. Conceitual I), a geração da representação do sumário correspondente (repres. Conceitual II) e a sua síntese, ou realização lingüística, resultando no *abstract*, propriamente dito[35]. Essa última etapa é responsável pelas escolhas morfossintáticas da língua natural em foco, as quais não necessariamente coincidem com as apresentadas no texto-fonte.

Segundo essa arquitetura, um sumarizador automático contempla três tipos de informação: o *lingüístico*, o *informativo* (ou de domínio) e o *comunicativo*, remetendo a questões semânticas e pragmáticas que aumentam a complexidade dos sistemas, devido à necessidade de modelagem do conhecimento necessário para manipulá-la. É necessário haver uma linguagem de representação que possibilite o inter-relacionamento entre as unidades de significado (aqui chamadas de *proposições*) e engenhos de inferência capazes de interpretar o texto-fonte e gerar sua forma condensada correspondente. Como são usados métodos simbólicos e modelos computacionais de geração textual bastante complexos para a manipulação de conhecimento profundo, é possível mesmo que *abstracts* contenham informações que não se encontram no texto-fonte, decorrentes de processos inferenciais sobre o conhecimento explícito no texto-fonte. Modelos para distinguir os diferentes graus de importância das informações dependem da caracterização dos interesses do escritor, os quais são regidos por objetivos comunicativos, e de modelos de estruturação do discurso, razão pela qual alguns métodos são também conhecidos como *métodos baseados em estruturação de discurso*.

[35] Lembrando que '*abstract*' é o termo que adotamos para diferenciar a metodologia empírica da fundamental.

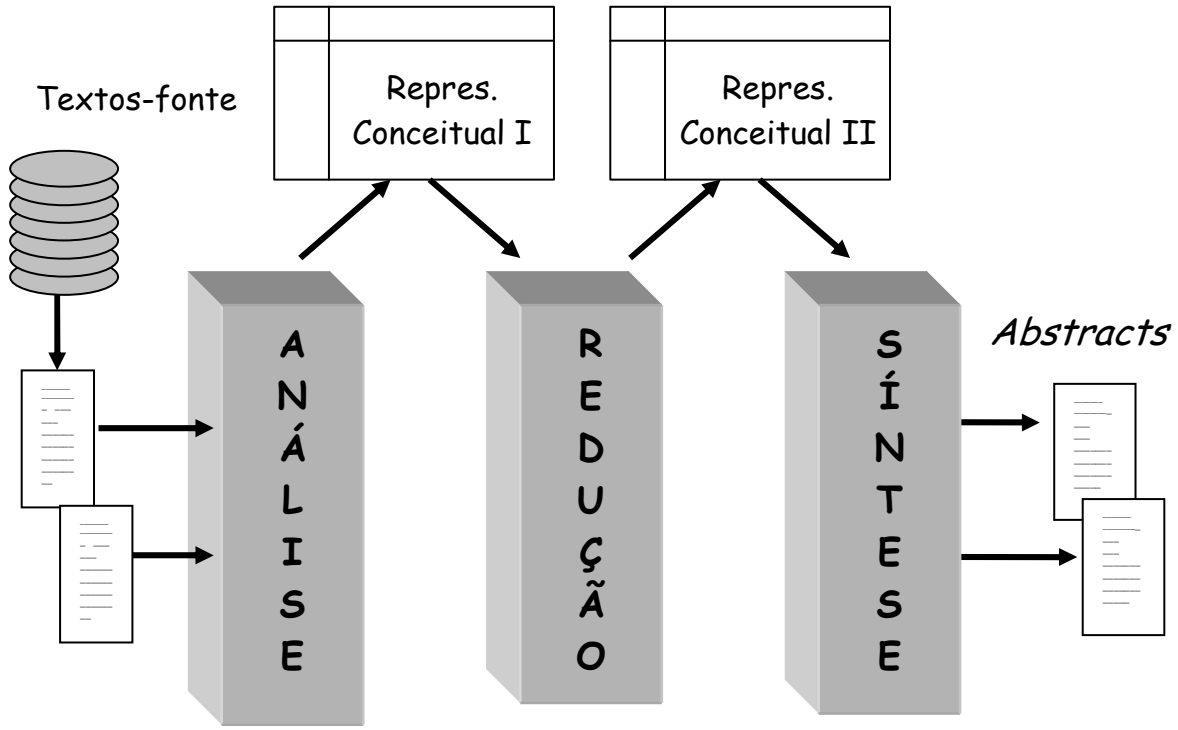


Figura 4. Sumarizador automático fundamental

O processo de análise, nessa abordagem, geralmente incorpora a um *parser* também um analisador discursivo, para produzir uma representação conceitual discursiva (e não sintática), pela qual se possam recuperar as relações entre os diversos segmentos textuais, assim como sua significância para a SA. Devido à natureza da estrutura global, são as proposições (e não suas expressões superficiais na forma de sentenças) que regem a análise. A premissa fundamental, neste caso, é que a idéia representada por um conjunto de proposições é estruturada de forma coerente antes mesmo das escolhas de vocabulário ou estrutura sintática, características intrínsecas da língua natural e não da linguagem do pensamento. Essa premissa justifica a coerência e coesão entre as unidades informativas, assim como a existência de diversos textos para uma mesma mensagem (diversas realizações lingüísticas para uma mesma estrutura conceitual).

Várias são as perspectivas dessa abordagem, para determinar as informações relevantes a partir da modelagem discursiva. A *saliência* das informações de um texto-fonte é uma propriedade importante, definida como a "*medida de proeminência relativa dos objetos ou conceitos textuais*" (Boguraev e Kennedy, 1997): as unidades informativas com grande saliência são o foco de atenção no discurso e, logo, devem ser consideradas nos *abstracts*; as com baixa saliência são periféricas e, logo, são passíveis de exclusão dos mesmos. Essa noção equivale á noção de significância ou relevância amplamente explorada na SA, que rege os critérios de escolha e agregação de segmentos textuais ou proposicionais, para a produção de sumários, de um modo geral.

O trabalho mais importante, hoje, nessa linha, é o de Marcu (1997a, 1997b, 2000), que propõe técnicas de segmentação do discurso para identificar o tópico e, a partir deste, estabelecer a saliência das informações relacionadas. A determinação das informações salientes é feita com base na estrutura retórica do texto, formalizada segundo a Teoria RST – *Rhetorical Structure Theory* (Mann e Thompson, 1988). Assim, é preciso primeiro construir a estrutura retórica do texto-fonte (tarefa de análise discursiva), para, então, determinar o conteúdo e a forma de seus possíveis sumários (tarefa de redução), ou seja, produzir a estrutura retórica do sumário correspondente. A vantagem dessa abordagem está na própria

definição das relações retóricas: elas indicam a assimetria do relacionamento proposicional, pela identificação de funções discursivas distintas. Estas, por sua vez, são construídas pela agregação de informações nucleares (os *núcleos*) e complementares (os *satélites*). Assim, Marcu explora a própria *nuclearidade* da Teoria RST, para identificar informações extraídas dos textos-fonte com diferentes graus de saliência.

O cômputo da saliência dos componentes do discurso se baseia tanto na nuclearidade quanto em sua profundidade na estrutura RST: núcleos mais próximos da raiz são considerados mais importantes do que seus satélites ou outros núcleos mais distantes da mesma. Uma possível estrutura RST para o Texto 1 é ilustrada na Figura 5 (N indicando núcelo e S o satélite). Cada proposição, neste caso, é delimitada por um segmento textual (numerado no Texto 1), sob a hipótese de que ele é a expressão superficial da proposição. As folhas da estrutura indicam, assim, as proposições, enquanto seus nós intermediários remetem às relações RST. Na Figura 5 são usadas somente as seguintes relações, com suas respectivas funções retóricas: *elaboration* (S elabora sobre N, apresentando detalhes e exemplos), *list* (as proposições fazem parte de uma lista de itens comparáveis, segundo algum critério de similaridade), *justify* (S justifica N), *purpose* (N é iniciado para realização de S).

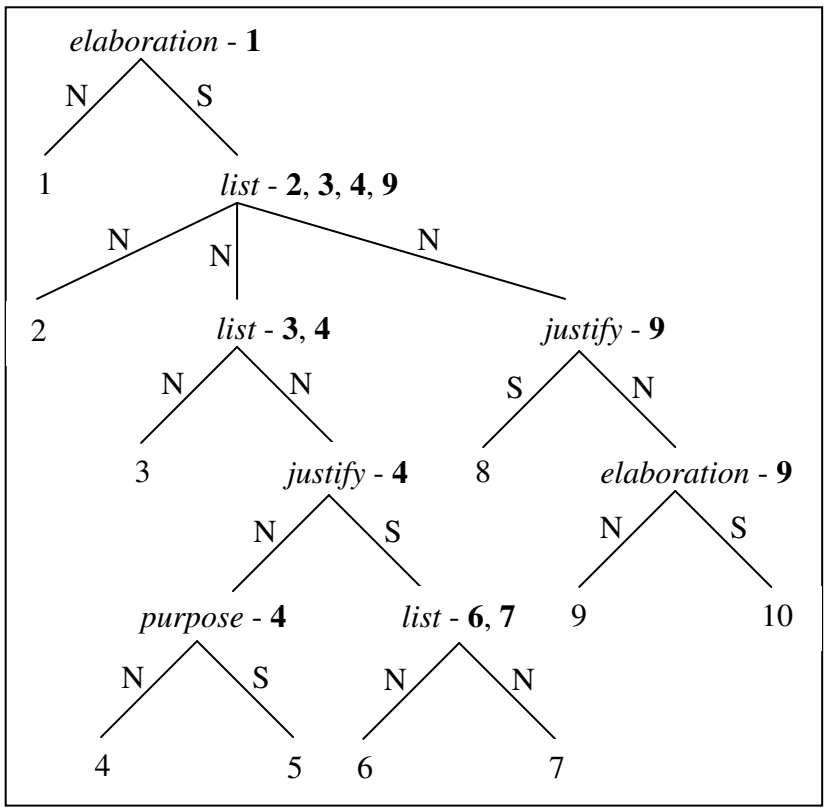


Figura 5. Estrutura RST do Texto 1

Nessa estrutura RST, as unidades mais salientes de cada segmento discursivo são indicadas junto aos nós intermediários. A ordem de precedência entre todas as proposições desse discurso é dada por 1>2>3>9>4>8>10>6=7>5 ('p1>p2' indica que p1 é mais importante que p2, assim como 'p1=p2' indica que p1 tem a mesma importância que p2). Sumários do Texto1 podem, agora, ser construídos respeitando-se essa ordem. Variando-se o número de segmentos a incluir, podemos ter as estruturas RST 1 e 2 da Figura 6, as quais podem ser expressas superficialmente, por exemplo, pelas manchetes M1 e M2, como sumários do Texto 1.

Vemos, assim, que a mensagem M1 envolve somente a relação *elaboration* entre 1 e 2; a mensagem M2 envolve também a relação *list* entre 2, 3, 4 e 9. Ambas, no entanto, têm a proposição 1 como mais saliente do discurso. Vale notar, também, que, por serem representações conceituais *profundas* da mensagem, essas mesmas estruturas poderiam derivar outras escolhas superficiais, produzindo textos diversos.

Essa proposta parece ser a mais consistente e efetiva atualmente, sendo independente de gênero e correlacionando-se à percepção que leitores têm sobre a importância de unidades textuais (Marcu, 1999). Entretanto, ela pressupõe a disponibilidade de estruturas RST para cada texto-fonte a sumarizar e, logo, requer um bom interpretador de língua natural, que gere suas estruturas retóricas. Atualmente, existem alguns interpretadores dessa natureza, para a língua inglesa (Marcu, 2000; Corston-Oliver, 1998; Schilder, 2002). Para o português, há uma proposta de análise discursiva em estágio inicial[36], associada ao modelo discursivo do sistema DMSumm, ilustrado na próxima seção.

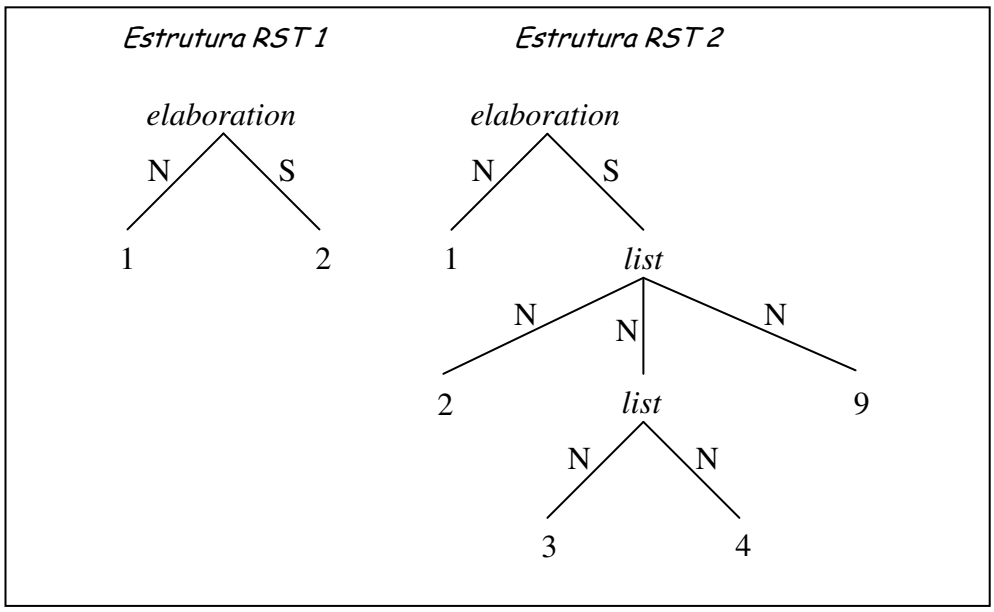


Figura 6. Possíveis estruturas RST das mensagens M1 e M2

14. Ilustrações

No NILC, exploramos tanto a abordagem empírica quanto a fundamental, sob o Projeto EXPLOSA[37]. Apresentamos, aqui, dois sistemas de cada abordagem (empírica e fundamental).

14.1. O GistSumm

O GistSumm, sigla para GIST SUMMarizer, é um sumarizador extrativo que faz uso de técnicas estatísticas simples para determinar a idéia central (o *gist*) dos textos a sumarizar. Ele procura simular a forma de sumarização humana, inicialmente identificando a idéia principal do texto e, então, acrescentando informações complementares. Assim, primeiramente ele

[36] Pardo, T.A.S. (2003). Análise discursiva automática de textos em português do Brasil. Proposta de Doutorado. ICMC/USP, Junho.

[37] Sigla para **EXPLO**ração de métodos de **S**umarização **A**utomática, Proc. FAPESP Nro. 01/08849-8 (http://www.dc.ufscar.br/~lucia/PROJECTS/EXPLOSA.htm) (último acesso em maio/2003).

procura a sentença que melhor expressa a idéia principal, chamada aqui de *gist sentence*, e, baseando-se nela, seleciona as demais sentenças para compor o extrato. Além das premissas básicas da SA, ele considera ainda que é sempre possível identificar, no texto-fonte, essa sentença. Com base nessas premissas, as hipóteses do GistSumm são as seguintes: (a) a identificação da *gist sentence* é possível com o uso de métodos estatísticos simples; (b) conhecendo-se a *gist sentence*, é possível produzir extratos coerentes por meio da justaposição de sentenças do texto-fonte relacionadas a ela. Consideramos que (a) pode ser confirmada também quando a sentença escolhida não for a *gist sentence*, mas uma aproximação significativa da mesma.

A exemplo dos métodos empíricos sem treinamento, o GistSumm compreende três processos: segmentação textual, ranqueamento e seleção de sentenças. A segmentação textual simplesmente delimita as sentenças do texto-fonte, procurando pelos sinais tradicionais de pontuação. Para o português, por exemplo, esses sinais incluem o ponto final e os sinais de exclamação e interrogação. O Texto 1 da Figura 1 é segmentado pelo GistSumm (sentenças numeradas). O ranqueamento consiste de sua ordenação, a partir de seus pesos, obtidos pela aplicação de métodos estatísticos. Ele ocorre em várias etapas, sendo que várias fases se aplicam a cada sentença. Para a sentença [7] do Texto 1, por exemplo – "O tipo mais grave da doença, o hemorrágico, pode matar." – os seguintes dados são manipulados em cada fase:

1) Vetorização das sentenças: Cada sentença é representada como um vetor (segundo Salton, 1988) cujas posições armazenam suas palavras.

O	tipo	mais	grave	da	doença	o	hemorrágico	pode	matar

2) *Case folding*, troca por canônicas e remoção de *stopwords*

Os processos de *case folding*, troca por canônicas e remoção de *stopwords* (sugeridos por Witten et al., 1994) são aplicados ao vetor de palavras. O *case folding* consiste em deixar todas as letras das palavras na mesma caixa (maiúscula ou minúscula) (por exemplo, a palavra "O" é trocada por "o"); a troca por canônicas simplesmente recupera do léxico do sistema (Nunes et al., 1996) a forma básica das palavras (por exemplo, a palavra "da" é trocada por "do"); a remoção de *stopwords* consiste em ignorar as palavras consideradas irrelevantes. Na sentença do exemplo, as canônicas das palavras "mais", "da" e "o" serão eliminadas. Sua remoção é realizada em três passos: (a) as palavras iguais são unificadas em uma única posição do vetor – a posição da primeira ocorrência da palavra; (b) a freqüência de cada palavra no vetor é armazenada junto às próprias palavras; (c) a freqüência das *stopwords* é zerada. Todos esses processos, além de facilitar o processamento computacional posterior, aprimoram os resultados da sumarização. Os vetores atualizados por cada processo, a partir do vetor inicial, são mostrados abaixo:

Case folding:

o	tipo	mais	grave	da	doença	o	hemorrágico	pode	matar

Troca por canônicas:

o	tipo	mais	grave	do	doença	o	hemorrágico	poder	matar

Remoção de *stopwords*:

o	tipo	mais	grave	do	doença	hemorrágico	poder	matar
0	1	0	1	0	1	1	1	1

3) Pontuação das sentenças

No GistSumm, a pontuação das sentenças pode ocorrer pelo uso de um dos seguintes métodos: palavras-chave (Black e Johnson, 1988) ou TF-ISF (*Term Frequency-Inverse Sentence Frequency*) (Larocca Neto et al., 2000). O método das palavras-chave segue a proposta de Luhn (1958), partindo do pressuposto de que a idéia principal de um texto pode ser expressa por um conjunto de palavras, chamadas chave. O método TF-ISF[38], por sua vez, determina a importância das sentenças de um texto, para escolher aquela que melhor o represente (a mais importante, no caso).

Em geral, os extratos produzidos por esses métodos são diferentes porque eles ponderam as sentenças de forma diversa. Pelo método das palavras-chave, cada vetor recebe como pontuação a soma do número de ocorrências de cada uma de suas palavras no texto inteiro (ou seja, em todos os vetores). No vetor anterior, há somente 1 palavra, no texto todo, com as seguintes canônicas: 'tipo', 'grave', 'hemorrágico', 'poder' e 'matar'. Há também 4 palavras com a canônica 'doença'. Assim, a pontuação total da sentença [7] é 5*1+ 1*4 = 9 (X*Y: X = número de canônicas; Y = número de palavras que remetem a uma única canônica).

Diferentemente desse cálculo, pelo método TF-ISF a pontuação do vetor corresponde à média da pontuação de cada uma de suas palavras. Sendo w uma palavra, essa pontuação é calculada da seguinte forma:

$$Pontuação\ de\ w = Fw \times \log\left(\frac{nro.\ palavras\ da\ sentença}{nro.\ sentenças\ em\ que\ w\ ocorreu}\right)$$

em que *Fw* é a freqüência da palavra na sentença. Para a sentença do exemplo, a pontuação obtida é de 0,689.

Por qualquer um dos métodos, a *gist sentence* do Texto 1 será a sentença com maior pontuação, como já mencionado. Por isso, no GistSumm os métodos de pontuação das sentenças são, na realidade, métodos de *determinação da idéia principal*. Para o Texto 1, coincidentemente a sentença [3] é escolhida como *gist sentence* por ambos os métodos. Essa sentença será sempre incluída no extrato, juntamente com as sentenças selecionadas por critérios de relevância e de taxa de compressão, as quais irão complementar a idéia principal extraída do texto-fonte. Esse processo de seleção é regido pelos seguintes passos:

1) calcula-se a média da pontuação das sentenças do texto-fonte e assume-se essa média como sendo um *cutoff*, nota de corte para eliminar sentenças irrelevantes do texto-fonte;
2) identificação das sentenças que contenham pelo menos uma palavra cuja canônica coincida com uma das canônicas da *gist sentence*;
3) dentre essas, seleção das que possuam uma pontuação maior que o *cutoff.*

Além disso, para respeitar a taxa de compressão especificada pelo usuário do sistema o GistSumm pode eliminar desse conjunto as sentenças com menor pontuação. As Figuras 7 e 8 apresentam os extratos produzidos pelo GistSumm pelo método das palavras-chave e pelo método TF-ISF, respectivamente, com uma taxa de compressão de 60%.

[38] O método TF-ISF é uma variação do método TF-IDF (*Text Frequency-Inverse Document Frequency*) (Salton, 1988) usado na área de Recuperação da Informação.

Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus.

Figura 7. Extrato produzido pelo GistSumm para o Texto 1 utilizando palavras-chave

Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. Assim, aumentaria a população de insetos refratários ao vírus.

Figura 8. Extrato produzido pelo GistSumm para o Texto 1 utilizando TF-ISF

Por fazer uso de métodos estatísticos, o GistSumm pode ser aplicado praticamente para textos de qualquer gênero, domínio ou língua ocidental, desde que se personalizem para a língua desejada seus repositórios lingüísticos, ou seja, o léxico e o repositório de *stopwords*. Mais detalhes sobre o sistema podem ser encontrados em (Pardo, 2002a; Pardo et al., 2003).

14.2. O NeuralSumm

O NeuralSumm, sigla para *NEURAL network for SUMMarization,* é um sumarizador extrativo que utiliza uma técnica de Aprendizado de Máquina – uma rede neural do tipo SOM (*self-organizing map*) (Braga et al., 2000) – para identificar as sentenças importantes de um texto-fonte. A classificação das sentenças em graus de importância é feita pela rede neural com base em *features* extraídas das sentenças durante o processo de sumarização.

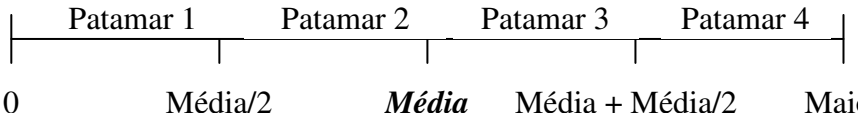
Diferentemente das outras redes neurais, uma rede do tipo SOM organiza as informações aprendidas em *clusters* (grupos) de similaridade. Justamente por isso, diz-se que esse tipo de rede é o que mais se aproxima da forma de funcionamento do cérebro humano. No NeuralSumm, as sentenças de um texto-fonte recebem sua classificação (o grau de importância) de acordo com os *clusters* da rede em que se enquadram. Em linhas gerais, o NeuralSumm extrai as *features* (descritas abaixo) de cada sentença do texto-fonte e as apresenta à rede neural, que as enquadrará em um dos *clusters* de similaridade, resultando na classificação da sentença com o valor associado a esse *cluster*.

É importante dizer que, por trabalhar com uma técnica de Aprendizado de Máquina, o NeuralSumm consiste em um sumarizador altamente experimental, pois se pode aumentar ou diminuir a rede (aumentando ou diminuindo o número de neurônios), alterar a arquitetura da rede, variar o conjunto de *features* utilizado, aumentar ou diminuir o número de *clusters* da rede, aumentar ou diminuir o tamanho do conjunto de treinamento e aumentar ou diminuir sua taxa de aprendizado e sua precisão à vontade para que se atinja a "melhor configuração" da rede, visando obter os melhores extratos.

A rede do NeuralSumm foi configurada para três *clusters*, representando as classes de sentenças *essenciais*, *complementares* e *supérfluas*, segundo as premissas básicas da SA antes delineadas. No estágio atual, ela é resultante de um treino com sentenças de um corpus de 10 textos científicos (introduções de teses e dissertações com aproximadamente 530 palavras e

19 sentenças cada, sobre Computação) em português, chamado *CorpusDT*[39] (Feltrim et al., 2001). Primeiramente, as sentenças desses textos foram classificadas de acordo com sua importância (valores *essencial*, *complementar* ou *supérflua)* por 10 juízes lingüistas computacionais e falantes nativos do português. Para cada uma delas, extraiu-se um conjunto de 8 *features*, associando a ele a classificação indicada pelos juízes. Este procedimento replica vários dos descritos para a abordagem extrativa. As *features*, assim como suas premissas e valores são reproduzidos abaixo:

1. Tamanho da sentença: sentenças longas normalmente apresentam maior conteúdo informativo, sendo, portanto, relevantes para o texto (Kupiec et al., 1995). No NeuralSumm, as sentenças de um texto são enquadradas em uma escala de 4 patamares. Considerando a variável *Média* como a média do tamanho das sentenças do texto em um intervalo de 0 a *Maior* (esta representando o máximo comprimento de sentenças no texto), essa escala é definida como segue:



2. Posição da sentença: a posição da sentença no texto pode indicar sua relevância (Baxendale, 1958)[40]. Seguindo Kupiec et al. (1995), no NeuralSumm uma sentença pode estar no *início* (primeiro parágrafo), no *fim* (último parágrafo) ou no *meio* (parágrafos restantes) do texto.
3. Posição da sentença no parágrafo a que pertence: a posição da sentença no parágrafo também pode indicar sua relevância (Baxendale, 1958). Da mesma forma que anteriormente, no NeuralSumm uma sentença pode estar no *início* (primeira sentença), no *fim* (última sentença) ou no *meio* (posições restantes) do parágrafo.
4. Presença de palavras-chave na sentença: as palavras-chave são comumente utilizadas para expressar a idéia principal do texto, tendendo a se repetir no decorrer do texto (Luhn, 1958). Assim, uma sentença pode conter (*True*) ou não (*False*) palavras-chave do texto. No NeuralSumm, elas são as palavras significativas (de classe aberta) de mais alta freqüência.
5. Presença de palavras da *gist sentence* na sentença: sentenças que possuem palavras da *gist sentence* tendem a ser mais relevantes do que outras, pois fazem alusão explícita à idéia principal do texto (Pardo, 2002b). Uma sentença pode conter (*True*) ou não (*False*) palavras da *gist sentence*.
6. Pontuação da sentença com base na distribuição das palavras do texto: sentenças com alta pontuação normalmente são relevantes para o texto (Black e Johnson, 1988). A pontuação de uma sentença, neste trabalho, é resultante da distribuição de suas palavras, isto é, da divisão da soma das freqüências de suas palavras por seu comprimento (número de palavras)[41]. Essa pontuação também é enquadrada em uma escala similar à da *feature* 1, cujos patamares (*1*, *2*, *3* ou *4*) indicam a representatividade da sentença.
7. TF-ISF da sentença: sentenças com alto valor de TF-ISF (*Term Frequency-Inverse Sentence Frequency*) são sentenças singulares de um texto e, assim, podem representá-lo bem (Larocca Neto et al., 2000). O valor TF-ISF de cada sentença também é enquadrado em uma escala similar à da *feature* 1, nos patamares *1*, *2*, *3* ou *4*.
8. Presença de palavras indicativas na sentença: palavras sinalizadoras (*cue words*) normalmente indicam a importância do conteúdo das sentenças (Edmundson, 1969; Paice,

[39] Descrição disponível em http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm (último acesso em maio/2003).
[40] Também confirmada por Aretoulaki (1996).
[41] Vale notar que, neste sistema, essa pontuação é distinta daquela do GistSumm.

1981). Uma sentença pode conter (*True*) ou não (*False*) palavras sinalizadoras. Essa *feature* é a única dependente de língua, gênero e domínio textuais. No NeuralSumm, atualmente customizado para textos científicos em português, as palavras sinalizadoras consideradas são *avaliação*, *conclusão*, *método*, *objetivo*, *problema*, *propósito*, *resultado*, *situação* e *solução*.

Após o treinamento da rede para esses conjuntos de *features* extraídos das sentenças do corpus de treino, os *clusters* delineados são usados para produzir extratos de textos-fonte, segundo os seguintes passos:

1. segmentação do texto-fonte em sentenças;
2. tratamento dos segmentos (remoção de *stopwords*; troca por canônicas; *case folding*);
3. extração das *features* de cada sentenças (a partir da representação interna produzida pelos passos anteriores);
4. classificação do conjunto de *features* de cada sentença segundo os *clusters* relativos aos valores *essencial*, *complementar* ou *supérfluo*;
5. seleção das sentenças com maior classificação e produção do extrato.

A seleção de sentenças para a produção do extrato (passo 5) acontece da seguinte forma:

- são selecionadas somente sentenças classificadas como *essenciais* e *complementares*;
- caso todas as sentenças do texto-fonte sejam classificadas como *supérfluas*, elas são ranqueadas pela pontuação obtida por sua distribuição de palavras (*feature* 6), selecionando-se, então, aquelas com pontuação mais alta.

Ambos os casos são ainda condicionados à taxa de compressão especificada pelo usuário do sistema. No primeiro caso, quando a taxa de compressão restringe o número de sentenças selecionadas, as sentenças *essenciais* têm prioridade sobre as *complementares*, sendo que sempre têm prioridade as de maior pontuação (pela *feature* 6).

As *features* extraídas da sentença [6] do Texto 1 (“A dengue provoca náuseas e dores de cabeça, articulações e músculos”), após a classificação de suas sentenças, são as seguintes:

1. tamanho da sentença: patamar 2 (11 palavras)
2. posição da sentença no texto: fim
3. posição da sentença no parágrafo a que pertence: meio
4. presença de palavras-chave: false
5. presença de palavras da *gist sentence*: true
6. pontuação da sentença com base na distribuição de palavras: patamar 2 (pontuação=1,428)
7. TF-ISF da sentença: patamar 3 (TF-ISF=0,731)
8. presença de palavras indicativas: false

Ao ser apresentado à rede neural, esse conjunto de *features* é enquadrado no *cluster* das sentenças supérfluas, determinando a classificação da sentença [6] como supérflua. O extrato correspondente, com taxa de compressão de 60%, é mostrado na Figura 9. Como podemos notar, a sentença [6] não foi incluída no extrato, já que é supérflua.

> O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. O tipo mais grave da doença, o hemorrágico, pode matar. Cerca de 1250 municípios brasileiros, aproximadamente um em cada quatro, registraram casos de dengue.

Figura 9. Extrato produzido pelo NeuralSumm para o Texto 1

Podemos notar, ainda, que esse extrato é ruim, pois não preserva a idéia principal do texto (esta é identificada pela sentença [3]). Esse extrato sequer menciona a palavra "dengue", crucial nesse texto, conforme evidenciam as manchetes de exemplo. Atribuímos a esse desempenho ruim o fato de o NeuralSumm ter sido treinado com um corpus de textos científicos da Computação. Logo, ele não é adequado para sumarizar o Texto 1, de gênero jornalístico e domínio muito distinto do domínio de Computação.

Esse exemplo mostra que, apesar de o NeuralSumm incorporar um método de SA genérico o suficiente para ser aplicado a qualquer texto de qualquer gênero e domínio, é preciso treiná-lo com corpora específicos a cada alteração de gênero ou domínio. Similarmente ao GistSumm, ele também pode ser aplicado para qualquer língua ocidental, bastando que se personalizem seus repositórios lingüísticos, que incluem agora o léxico, o repositório de *stopwords* e o repositório de palavras indicativas.

14.3. O DMSumm

DMSumm é a sigla para *Discourse Modeling SUMMarizer*, um gerador automático de sumários (Pardo, 2002b; 2002c) baseado em modelagem discursiva (Rino, 1996). Embora vise à SA de textos, ele não tem como entrada o próprio texto, mas sim o suposto resultado de sua interpretação. Essa delimitação foi adotada devido à inexistência de um interpretador adequado para a modelagem do discurso considerada, à complexidade de se construir um e ao interesse mais imediato de se explorar as questões peculiares da SA. Assim, no momento, a entrada para o DMSumm é produzida manualmente.

A mensagem (ou representação conceitual, cf. nomenclatura da Figura 4) de um texto-fonte é composta por três componentes: uma base de conhecimento, uma proposição central e um objetivo comunicativo. A base de conhecimento é uma estrutura semântica que contém o conhecimento expresso no texto-fonte; a proposição central é a informação mais importante do texto, aquela que se quer comunicar, e o objetivo comunicativo é o objetivo que se quer atingir ao comunicar o conteúdo do texto-fonte. Com essa caracterização, o DMSumm observa as três premissas fundamentais da SA: todo texto tem um objetivo comunicativo e uma proposição central, devendo esta ser preservada na sumarização. A hipótese principal do DMSumm está na garantia de coerência dos sumários pela interação de conhecimento de diferentes naturezas – semântica, intencional e retórica – presentes na modelagem discursiva utilizada.

A base de conhecimento é uma árvore binária cujos nós internos são rotulados por relações semânticas (representadas em itálico na Figura 10), baseadas nas relações clausais de Jordan (1992), e cujas folhas são proposições correspondentes às unidades informativas do texto-fonte. Além das relações semânticas, também é registrado o papel funcional dos segmentos no texto-fonte. Para o Texto 1, por exemplo, cuja base de conhecimento é ilustrada na Figura 10, a relação *rationale* entre os segmentos 3 e 4 indica que, com a realização de sol(3), tem-se o objetivo de realizar prop(4). Os papéis funcionais dos segmentos são expressos de duas maneiras: como etiquetas associadas a cada segmento (sit=situação, probl=problema, sol=solução, res=resultado e prop=proposição genérica) e como blocos semânticos (Situação, Problema, Solução e Resultados) indicando o papel que um conjunto de segmentos desempenha no texto. Assim, podemos dizer que as primeiras etiquetas contemplam o nível micro-estrutural, informativo, do texto-fonte, enquanto que as últimas contemplam seu nível macro-estrutural[42]. sol(3), por exemplo, indica que o segmento 3 é uma

[42] Muito embora alguns autores associem a esse nível o conhecimento retórico do texto (como o fazem Teufel e Moens, 1999), nós seguimos sobretudo a linha de Winter (1977; 1979), em que as funções dos segmentos textuais não expressam, necessariamente, qualquer omposição de objetivos retóricos ou conotação.

solução para algum problema apontado no texto (que, no caso, é probl(2)) e que os segmentos 2, 6, 7, 8, 9 e 10 fazem parte da descrição do problema apresentado no texto. Essa base de conhecimento expressa todo o conteúdo disponível para a produção de um sumário, no DMSumm.

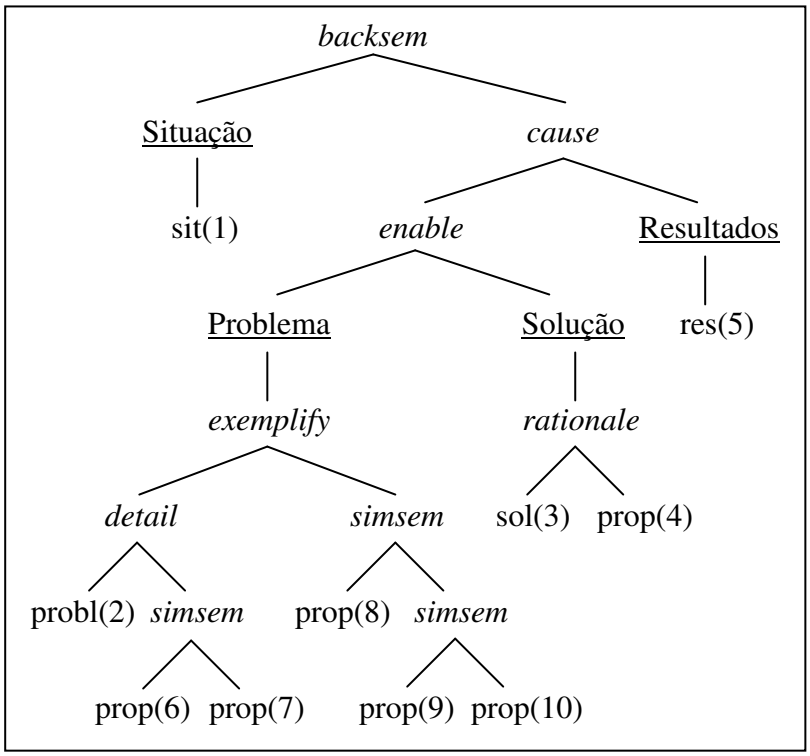


Figura 10. Base de conhecimento do Texto 1

Um possível objetivo comunicativo para o Texto 1 é *relatar a solução encontrada para o problema*. Desse objetivo, é possível, portanto, derivar a proposição central do possível sumário: a solução indicada pelo segmento 3. Ele também é responsável por delimitar a estratégia comunicativa para a construção do sumário. A proposição central, por sua vez, será a informação principal que restringirá a escolha de outros segmentos do texto-fonte, os quais deverão complementá-la, visando maior informatividade ou coerência do sumário. No DMSumm, consideram-se somente os objetivos comunicativos *descrever*, *relatar* e *discutir*.

Tendo os três componentes da mensagem do texto-fonte disponíveis como entrada (a base de conhecimento, o objetivo comunicativo e a proposição central), o DMSumm pode aplicar sua estratégia fundamental, de transformação e síntese dos possíveis sumários. São três os processos correspondentes a essas etapas de SA: a seleção de conteúdo, o planejamento textual e a realização lingüística. O processo de seleção de conteúdo simplesmente elimina da base de conhecimento suas proposições supérfluas, identificadas pela falta de relação expressiva com o objetivo comunicativo e a proposição central. São usadas heurísticas para essa redução de conteúdo (Rino e Scott, 1994). Por exemplo, ao excluir exemplos e detalhes da base de conhecimento do Texto 1 (relações semânticas *exemplify* e *detail*), a base de conhecimento reduzida resulta na ilustrada na Figura 11, cujas proposições devem, agora, ser reproduzidas no sumário final.

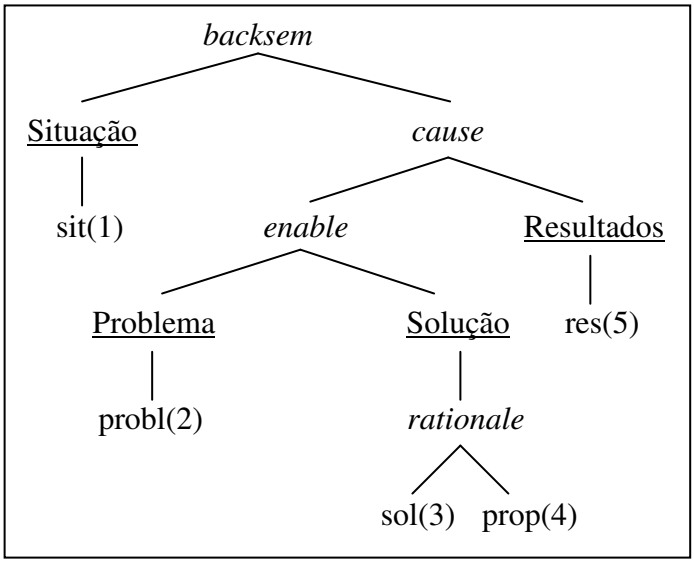


Figura 11. Base de conhecimento reduzida do Texto 1

O processo de planejamento textual organiza o conteúdo informativo restante, produzindo *planos de texto*, i.e., estruturas retóricas, por meio de um modelo de discurso (Rino, 1996) multi-nível: ele fundamenta a escolha de relações retóricas (Mann e Thompson, 1988) pelo mapeamento das relações semânticas indicadas na base de conhecimento e das relações intencionais (Grosz e Sidner, 1986) delineadas pelo objetivo comunicativo. As relações intencionais irão determinar a força retórica das unidades informativas, enquanto as semânticas irão delinear a forma como elas serão relacionados na estrutura final, isto é, se serão componentes de um núcleo ou de um satélite da estrutura retórica do sumário. Esse mapeamento é modelado computacionalmente por *operadores de plano* (Moore e Paris, 1993), artifícios que permitem identificar restrições e buscar a satisfação de condições para a determinação da estrutura e do conteúdo textual, garantindo a construção do plano de texto. Em seu estágio atual, o DMSumm incorpora 89 operadores de plano, responsáveis por gerar todos os mapeamentos possíveis entre as relações do modelo de discurso implementado. A Figura 12 apresenta um plano de texto para a mensagem antes ilustrada, de relatar o problema relacionado ao conteúdo da base de conhecimento (Figura 10) tendo como proposição central sua solução (*sol*(*3*)).

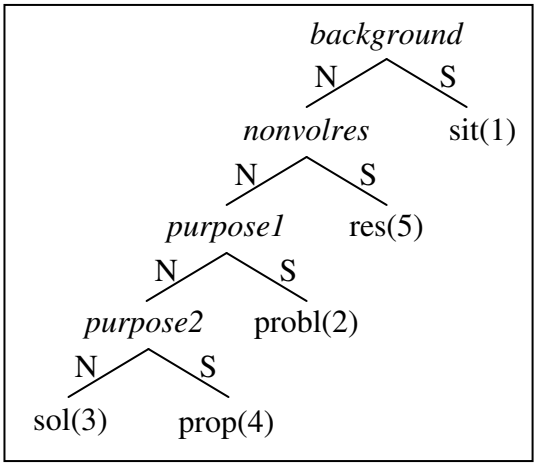


Figura 12. Plano de texto de um sumário do Texto 1

O último processo, de realização lingüística, é responsável por expressar em língua natural os planos de texto, produzindo, enfim, os sumários. Este processo é ainda simplificado no DMSumm, pois envolve somente o uso de *templates* para as escolhas de vocabulário e

sintaxe, definidos por um conjunto de regras fixas de expressão das relações retóricas entre as unidades informativas. Quando necessário, são incluídos marcadores discursivos e sinais de pontuação para garantir a formação adequada do sumário. O sumário apresentado na Figura 13 é um exemplo de realização lingüística do plano de texto da Figura 12.

Sumário 1: Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. Em face disso, em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. Como resultado, aumentaria a população de insetos refratários ao vírus.

Figura 13. Sumário produzido pelo DMSumm para o Texto 1

De forma similar à ilustrada, é possível gerar vários sumários para o Texto 1, variando-se o objetivo comunicativo e/ou a proposição central.

Por ser baseado em um modelo de discurso independente de língua, o DMSumm pode ser aplicado praticamente a qualquer língua, bastando que se personalize seu módulo de realização lingüística. É importante notar, também, que os textos a serem sumarizados pelo DMSumm devem apresentar uma estrutura a partir da qual possam ser derivados os componentes e as relações pertinentes ao modelo de representação da base de conhecimento, como os ilustrados na Figura 10. Ou seja, devem apresentar um problema e sua solução, assim como os resultados obtidos, etc. Esse tipo de estruturação segue, na realidade, o modelo Problema-Solução (Winter, 1976; 1977; Jordan, 1980, 1984; Hoey, 1983), que é bastante discutido na literatura e freqüentemente encontrado em textos de diversos gêneros e domínios, escritos em quaisquer línguas naturais.

Para mais detalhes sobre o DMSumm, incluindo a especificação de suas relações semânticas, intencionais e retóricas e a definição dos operadores de plano e *templates* utilizados, vide (Pardo, 2002b; 2002c) e (Pardo e Rino, 2002).

14.4. O UNLSumm

Diferentemente do DMSumm, o UNLSumm, ou *UNL SUMMarizer* (Martins, 2002), é um sumarizador sentencial implementado em plataforma Windows que, embora baseado em metodologia fundamental, não faz uso profundo de características discursivas ou retóricas, mas sim das características conceituais da Linguagem UNL (Uchida, 2000), a interlíngua adotada no Projeto UNL-Brasil, de tradução multilingual.

Sua entrada é uma estrutura UNL, representação conceitual, na Linguagem UNL, de cada sentença de um texto-fonte, o qual, em princípio, pode ser escrito em qualquer língua natural. Portanto, a fase de análise, nesse sistema, também é considerada independente do UNLSumm e, por essa razão, as representações conceituais são construídas ainda manualmente. A redução de uma estrutura UNL é fundamentada, assim, no inter-relacionamento semântico entre unidades de informação, formalizado pela Linguagem UNL. Supõe-se que a estrutura UNL reduzida poderá ser sintetizada com o auxílio do DeCo (Uchida, 1997), ferramenta de decodificação de UNL para qualquer língua natural desejada. Particularmente no contexto do UNLSumm, contemplamos somente o português, já que é para essa língua que o DeCo foi personalizado no Projeto UNL-Brasil. Assim, a única etapa

da SA fundamental que o UNLSumm realmente contempla, no momento, é a de redução da representação conceitual original, para a produção de uma representação conceitual do sumário a gerar. Esta última estrutura, ainda em linguagem UNL, é chamada de *sumário UNL*.

A linguagem UNL é utilizada, assim, como linguagem de representação do conhecimento e fonte para o modelo de SA: o UNLSumm é baseado em conjuntos de heurísticas que, pela identificação de relações UNL, indicam os componentes menos relevantes da estrutura UNL, que serão, portanto, excluídos do sumário UNL. Diferentemente das demais propostas já apresentadas neste capítulo, este sistema concentra suas decisões em mecanismos de *exclusão*: são identificadas e extraídas as informações irrelevantes da estrutura UNL de entrada.

Para a especificação das heurísticas de identificação dos componentes supérfluos, a modelagem profunda consistiu da identificação das principais características conceituais de textos a sumarizar: foram considerados vários corpora de textos em português, para cujas sentenças foram produzidas manualmente suas estruturas UNL. Comparando a forma superficial com a conceitual, identificaram-se as correspondências lingüístico-conceituais entre construções superficiais do português e os componentes da Linguagem UNL, sobretudo aqueles remetendo aos rótulos de relação, ou *Relation Labels* (RLs), pois são estes que indicam o relacionamento conceitual entre diferentes componentes sentenciais. Estes, por sua vez, são representados em UNL pelas *Universal Words* (UWs).

O dado de entrada do UNLSumm, i.e., uma estrutura UNL, é expresso por um conjunto de relações binárias, semânticas, entre os componentes sentenciais, cujo formato é *RL(UW1,UW2)* – RL é uma relação conceitual entre dois conceitos distintos, sendo que estes podem ser simples ou compostos. Os RLs são expressos por uma cadeia de três caracteres. Por exemplo, *agt* é um RL que indica o agente de uma ação, como em "João quebra a janela da sala.", cuja relação binária é *agt(break, Joao)*.

As premissas básicas do sumarizador sentencial são as seguintes:

1. Sentenças em *pipeline* podem ser consideradas, quando conjugadas, um texto completo.
2. É possível produzir, para um texto-fonte, um sumário com mesmo número de sentenças, porém, com estruturas correspondentes condensadas, também considerando a produção de estruturas superficiais em *pipeline*, sentença por sentença. Assim, sucessivas sumarizações intra-sentenciais resultarão em representações UNL bem formadas.
3. Ao decodificar uma estrutura UNL em textos em certa língua natural, o processamento seqüencial de suas sentenças UNL ainda garantirá textos bem formados.

Por tratar somente da SA intra-sentencial, o UNLSumm impede que sejam usadas taxas de compressão variadas para a produção dos sumários. Embora não haja um consenso sobre o tamanho ideal de um sumário, normalmente considera-se que bons sumários mantenham de 5 a 30% do conteúdo do texto-fonte (Mani, 2001), ou seja, suas taxas de compressão variam de 70% a 95%. Para o UNLSumm, essas taxas são muito altas, devido ao fato de não se excluirem sentenças quaisquer da estrutura de entrada. Por essa razão, são considerados úteis também os sumários com baixa taxa de compressão.

Um exemplo do UNLSumm em operação é ilustrado na SA da sentença [3] do Texto 1 – “Em estudo publicado na edição de hoje da revista científica Science, pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus.”. Sua estrutura UNL completa indica várias relações binárias, como ilustra a Figura 14.

```
[S]
obj(published,study.@indef)                              [1]
plc(published,edition.@def)                              [2]
tim(edition.@def,today)                                  [3]
obj(edition.@def,magazine.@def)                          [4]
nam(magazine.@def,Science)                               [5]
mod(magazine.@def,scientific)                            [6]
plc(researcher.@pl,study.@indef)                         [7]
src(researcher.@pl,State University of Colorado)         [8]
plc(State University of Colorado,USA.@parenthesis)       [9]
agt(researcher.@pl,create.@entry.@past)                  [10]
obj(create.@entry.@past,mosquito.@indef)                 [11]
pos(organism,mosquito.@indef)                            [12]
aoj(accept.@not,organism)                                [13]
obj(accept.@not,carry)                                   [14]
obj(carry,virus.@def)                                    [15]
[/S]
```

Figura 14. Estrutura UNL da sentença [3]

O UNLSumm utiliza dois grupos de heurísticas para identificar as relações binárias indicativas de informações irrelevantes, em um total de 84 heurísticas. As heurísticas do primeiro grupo (Grupo A) identificam relações binárias que possam ser individualmente removidas da sentença original, enquanto as do segundo grupo (Grupo B) identificam grupos de relações binárias que devem ser removidas simultaneamente, não só por sua irrelevância, mas, principalmente, para garantir a coerência e coesão da sentença sumarizada. Como já mencionamos, as heurísticas se baseiam nos RLs. Por exemplo, para a sentença [3], são aplicadas duas heurísticas do grupo B, gerando o sumário UNL da Figura 15. Se este fosse decodificado novamente para o português, equivaleria à sentença "Em estudo publicado, pesquisadores criaram um mosquito cujo organismo não aceita carregar o vírus.".

```
[S]
obj(published,study.@indef)                              [1]
plc(researcher.@pl,study.@indef)                         [7]
agt(researcher.@pl,create.@entry.@past)                  [10]
obj(create.@entry.@past,mosquito.@indef)                 [11]
pos(organism,mosquito.@indef)                            [12]
aoj(accept.@not,organism)                                [13]
obj(accept.@not,carry)                                   [14]
obj(carry,virus.@def)                                    [15]
[/S]
```

Figura 15. Estrutura UNL de um sumário da sentença [3]

As duas heurísticas aplicadas, nesse caso, são as seguintes:

Heurística HB.3.5

Excluir plc(a,b) + {RBs ∈ subgrupo S1} se UWs ∈ S1 ∉ RBs fora do subgrupo.

Heurística HB.7.1

Excluir src(a,b) + {RBs ∈ subgrupo S1} se UWs ∈ S1 ∉ RBs fora do subgrupo.

A heurística HB.3.5 parte do princípio de que a informação sobre o local onde uma ação ocorre não é essencial e, portanto, pode ser omitida em um sumário. Como em UNL esse tipo de informação é representada por meio do RL plc (*place*, ou lugar), então a relação binária que contém esse RL deve ser excluída. Assim, a relação binária [2], correspondendo à informação "a edição", deve ser excluída da estrutura UNL. Sua exclusão caracteriza o subgrupo composto pelas relações binárias [2], [3], [4], [5] e [6], as quais envolvem as UWs *edition* e *magazine* (mais refinadas na estrutura ilustrada): essas UWs não aparecem no restante da estrutura UNL. Desse modo, a aplicação integral da heurística implica excluir também tais relações.

Já a heurística HB.7.1 parte do princípio de que a informação sobre a origem de um objeto conceitual (no exemplo, "Universidade Estadual do Colorado") não é essencial e, portanto, também deve ser omitida. Como em UNL esse tipo de informação é representada por meio do RL src (*source*, ou fonte), essa heurística exclui a relação binária envolvendo esse RL, [8], assim como aquela caracterizada no mesmo subgrupo – a de número [9].

15. A avaliação de sumários produzidos automaticamente

Esta seção apresenta uma visão geral sobre a questão da avaliação de sistemas de SA. A subseção seguinte introduz o tema, mostrando sua necessidade e as dificuldades associadas. A seguir, relatam-se definições e princípios gerais de avaliação adotados em pesquisas recentes, enquanto as subseções posteriores descrevem métricas e técnicas de avaliação comumente utilizadas. A última subseção mostra um estudo de caso, descrevendo a avaliação do GistSumm.

15.1. A necessidade e as dificuldades da avaliação

A avaliação de sistemas de PLN, em especial de sistemas de SA, foi bastante negligenciada no passado, mas muito importante ultimamente. É por meio da avaliação que se torna possível verificar o avanço do "estado da arte" em AS. É possível medir-se o grau de utilidade de um sistema de SA, sua adequação a determinadas tarefas, a validade de sua metodologia, etc. Em SA, pode-se dizer que a avaliação é a responsável por direcionar as pesquisas, pois ela pode indicar meios de validação e mesmo de desconsideração de orientações antes delineadas.

Esse tema é muito amplo e abrangente. Quando se fala em avaliação de um sistema, pode-se ter em mente várias facetas: (a) pode-se avaliar o desempenho computacional do sistema, isto é, o uso que ele faz de memória, seu tempo de execução, a complexidade de seu algoritmo principal, etc.; (b) pode-se considerar a usabilidade do sistema, ou seja, a crítica da clareza de sua interface ou o grau de intuição (dos possíveis usuários) necessário para seu uso ou, ainda, sua consistência e sua flexibilidade para possíveis customizações; (c) pode-se avaliar se os resultados produzidos automaticamente são satisfatórios, isto é, se são os resultados esperados, se são "corretos" ou "adequados". Na SA, contempla-se, em geral, somente esta última forma de avaliação, pois não há métodos de SA suficientemente robustos que justifiquem a análise de questões de desempenho ou interface.

Na literatura básica de SA, há muitas referências sobre métricas e métodos de avaliação. Porém, ainda não há consenso sobre a melhor forma de se avaliar um sistema dessa

natureza. Há, ainda, diversos desafios nessa área, dentre os quais destacam-se (Mani e Maybury, 1999):

- A identificação do que seria um resultado "correto" para um sumarizador automático, já que, para um único texto-fonte, pode haver uma infinidade de sumários sugerindo diversas perspectivas, em função de todos os possíveis usuários e das tarefas a que podem se destinar. Por exemplo, para um leitor leigo no assunto de um certo texto-fonte, toda informação contextual pode ser importante, enquanto que, para um leitor especialista, somente a informação nova poderia lhe interessar e, assim, ser incluída no sumário. Para um leitor decidir se lerá ou não o texto-fonte correspondente a partir da leitura de um sumário, este deve ser suficientemente informativo. Entretanto, se ele servir somente para indexar documentos, ele pode ser simplesmente uma lista de palavras. Mesmo para leitores de mesmo perfil e com tarefas comuns, um sumário pode ser julgado adequado por alguns e inadequado por outros.
- A identificação de uma taxa de compressão ideal para avaliar adequadamente os sumários automáticos. Em geral, quanto mais alta a taxa de compressão, menos informativo será o sumário e vice-versa. Entretanto, essa relação de dependência não pode ser explicitamente associada a um modelo fixo, pois a informatividade depende do nível de conhecimento do usuário ao qual o sumário se destina, do tempo que ele dispõe para lê-lo e da tarefa especificada, dentre outras coisas.
- A forma como a qualidade e a informatividade de sumários automáticos podem ser avaliadas automaticamente. Nesses casos, costuma-se fazer uso do julgamento humano: leitores falantes da língua natural considerada devem dizer se os sumários automáticos são bons sumários, quando comparados a seus correspondentes textos-fonte. Essa tática torna o processo de avaliação bastante custoso, tornando preferível uma avaliação automática. Entretanto, não há nada até o momento que seja capaz de substituir de forma satisfatória o juízo humano na avaliação.
- A identificação da situação e da forma de se utilizar o julgamento humano. Apesar de a resposta para essa questão parecer simples, inferindo-se que o julgamento humano deve ser utilizado em todas as fases possíveis de uma avaliação, há o problema de se identificar o perfil adequado do juiz humano, além do custo envolvido, já mencionado. A avaliação se torna custosa, pois: (a) é difícil dispor de juízes humanos (e, quando necessário, especialistas em técnicas de sumarização) em número suficiente para a avaliação; (b) para uma avaliação robusta e abrangente, esse tipo de julgamento se torna lento e complexo; (c) há um alto grau de subjetividade no julgamento humano, sendo difícil tirar conclusões definitivas com esse tipo de avaliação.

O surgimento de conferências internacionais dedicadas somente à avaliação de sumarizadores automáticos, como a SUMMAC[43] (*Text Summarization Evaluation Conference*) (Mani et al., 1998) e a DUC[44] (*Document Understanding Conference*), evidencia a importância e necessidade da avaliação e das dificuldades inerentes à AS. A SUMMAC foi realizada em 1998 e financiada pelo governo dos EUA, sendo a primeira avaliação independente, em larga escala, de sistemas de SA. Seu principal objetivo foi tentar estabelecer padrões de avaliação e entender melhor as questões envolvidas no processo integral de construção e avaliação de sistemas de SA. A DUC, por sua vez, nasceu de um programa chamado TIDES (*Translingual Information Detection, Extraction, and Summarization*) financiado pela DARPA (*Defense Advanced Research Projects Agency*), uma agência do Departamento de Defesa dos EUA,

[43] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/ (último acesso em maio/2003).
[44] http://duc.nist.gov/ (último acesso em maio/2003).

com os mesmos objetivos da SUMMAC. Entretanto, ela tem sido realizada periodicamente e é considerada a iniciativa atual mais importante de avaliação de sistemas de SA.

Apesar da problemática envolvida na avaliação de sistemas de SA, alguns princípios gerais e definições comuns têm sido adotados, conforme ilustra a próxima seção.

15.2. Avaliação: definições e princípios gerais

Sparck Jones e Galliers (1996) foram os primeiros autores a esclarecer as possíveis diretrizes gerais para a avaliação de sistemas de PLN, as quais têm sido amplamente adotadas.

A primeira grande distinção que se faz diz respeito à forma de avaliação: ela pode ser *intrínseca* ou *extrínseca*. Uma avaliação intrínseca avalia o próprio desempenho do sistema, pela verificação da qualidade e informatividade dos sumários produzidos. São usadas métricas calculadas automaticamente ou julgamentos subjetivos, realizados por leitores humanos. A avaliação extrínseca verifica a adequação do sistema ao seu uso em tarefas específicas, distintas da SA. Por essa razão, ela é comumente chamada de *validação*. Tarefas em que a validação de sumarizadores automáticos tem se aplicado envolvem as de perguntas e respostas, de categorização de documentos e de recuperação de informação.

Quando a avaliação faz uso do julgamento humano, diz-se que ela é uma avaliação *on-line*. Caso contrário, ela é chamada de *off-line*. Dada a complexidade de se projetar uma avaliação por seres humanos, a avaliação *off-line* é, normalmente, preferível. Entretanto, métodos automáticos de avaliação que sejam tão satisfatórios quanto o julgamento humano são ainda inexistentes.

Pode-se classificar a avaliação de acordo com o que se avalia: se forem avaliados somente os resultados finais do sistema, a avaliação é chamada *avaliação black-box*. Neste caso, o sistema é visto como uma "caixa-preta", à qual não se tem acesso. Ou seja, não se avaliam os processos intermediários da sumarização. Exemplo típico desse tipo de avaliação é a comparação entre um sumário produzido automaticamente e seu correspondente texto-fonte, para verificar se ele é bom. Se forem avaliados resultados intermediários, isto é, aqueles resultantes da execução de cada processo intermediário do sistema, a avaliação é chamada *avaliação glass-box*. Caso um sistema de SA siga a arquitetura padrão composta pelos processos de análise, transformação e síntese, uma avaliação glass-box verificaria os resultados de cada um desses processos.

Uma última distinção se faz em relação à forma de comparação entre vários sistemas. Se os resultados de um sistema de SA são comparados com os resultados de outro sistema, diz-se que a avaliação é *comparativa*; caso contrário, diz-se que ela é *autônoma*. A avaliação comparativa é, normalmente, o foco das grandes conferências internacionais (SUMMAC e DUC inclusas): os sistemas participantes do concurso são pontuados pelo seu desempenho e, então, são comparados pelos seus pesos.

Sparck Jones e Galliers afirmam que o mais importante na avaliação de um sistema de SA é estabelecer claramente o que se quer avaliar. Tendo isso como meta, é fácil determinar quais dos tipos anteriores de avaliação aplicar, isto é, se ela será intrínseca ou extrínseca, on-line ou off-line, black-box ou glass-box e comparativa ou autônoma. É importante esclarecer, entretanto, que estes tipos de avaliação não são exclusivos. Por exemplo, caso se queira proceder a uma avaliação intrínseca e a uma extrínseca, isso é totalmente possível e viável, dependendo somente dos objetivos ao se realizar a avaliação.

A próxima seção apresenta os métodos e métricas usuais para a avaliação intrínseca de sistemas de SA, seguindo-se aqueles da avaliação extrínseca.

15.3. Avaliação intrínseca

Em geral, a avaliação intrínseca pode envolver as medidas de *qualidade* e *informatividade* dos sumários produzidos automaticamente (Mani, 2001).

15.3.1. Qualidade dos sumários automáticos

Nos termos de Mani, medir a qualidade de sumários é verificar a sua *fluência*, ou seja, a facilidade em sua leitura e a sua clareza. Assim, é necessário o julgamento humano e, por essa razão, esse tipo de avaliação é geralmente caracterizado como avaliação *on-line.*

Os critérios para se julgar a qualidade dos sumários variam muito. Minel et al. (1997), por exemplo, pediram a juízes humanos que dessem notas a sumários observando os seguintes critérios: presença de referências anafóricas não resolvidas, não preservação da integridade de estruturas como listas e tabelas, falta de coesão entre as sentenças do sumário, presença de tautologias (que é um vício de linguagem que consiste em repetir o mesmo pensamento com palavras diferentes), etc. Saggion e Lapalme (2000), utilizando os critérios sugeridos por Rowley (1982), também pediram a juízes que dessem notas aos sumários, observando agora a ortografia e gramática, a indicação clara do tópico do texto-fonte, o estilo impessoal, a concisão, legibilidade e facilidade de compreensão do sumário, a presença de siglas seguidas de suas expansões, etc. Pardo e Rino (2002), seguindo os critérios de avaliação de White et al. (2000), também pediram a juízes que dessem notas a sumários de acordo com sua *textualidade*, isto é, sua coerência e coesão (Rino, 1996). Além disso, usaram também outra sugestão de Mani, de avaliação da legibilidade dos sumários, ante a legibilidade dos textos-fonte correspondentes. Para verificar se os sumários preservavam a legibilidade dos textos-fonte, foi utilizado o índice de legibilidade de Flesch (1948) adaptado ao português (Martins et al., 1996) e calculado automaticamente (*off-line*). Esse índice baseia-se no número médio de sílabas por palavra e de palavras por sentenças, para expressar o grau de dificuldade de leitura de um texto. É importante ressaltar que a legibilidade não é um critério decisivo, nem suficiente, para se afirmar que um sumário é bom. De fato, como discutido por Mani, essa medida é muito rústica, dada sua "ingenuidade" em assumir que o tamanho de palavras e de sentenças é o único fator que pode influenciar a legibilidade de um texto.

Como a avaliação da qualidade de sumários necessita de juízes humanos, tem-se procurado formas automáticas de realizar tal avaliação. A verificação da legibilidade dos sumários em relação aos textos-fonte é um exemplo de automação factível. Outras opções incluem o uso de corretores ortográficos, gramaticais ou estilísticos automáticos.

A qualidade é um bom parâmetro para se centralizar avaliações de sistemas extrativos, pois estes produzem, em geral, textos com fluência ruim. Entretanto, mesmo com fluência prejudicada, ainda é possível obter sumários úteis. Devido à complexidade de modelagem de tantas nuances distintas sobre a qualidade e/ou utilidade de sumários automáticos, tornou-se comum avaliá-los somente pela verificação do conteúdo que eles preservam, em relação a seus textos-fonte.

15.3.2. Informatividade dos sumários automáticos

A informatividade de um sumário expressa o quanto, do conteúdo informativo original, ele contém. Na maioria das aplicações, senão todas, esse é um quesito essencial para julgar a qualidade de um sumário.

A informatividade está diretamente ligada à taxa de compressão: é comum considerar-se que, quanto maior a compressão, menos informativo será o sumário e vice-versa, muito embora isto nem sempre seja verdadeiro, principalmente se levarmos em conta a competência de domínio do leitor, a qual poderá modificar significativamente os critérios de informatividade. Assim, considerando o caso de leitores de proficiência média, pode-se dizer que a redução expressiva do conteúdo do texto-fonte pode prejudicar sensivelmente a informatividade do sumário. Desta forma, torna-se necessário determinar a taxa de compressão adequada, para não haver prejuízos consideráveis sobre a informatividade dos sumários. Em geral, essa taxa pode ser inferida por observações: analisando tarefa(s) que usuário(s) executam com o uso de sumários, é possível caracterizá-los mais apropriadamente.

Para se verificar a informatividade de um sumário, além de comparar seu conteúdo com o conteúdo de seu texto-fonte, pode-se compará-lo também com o conteúdo de um sumário de referência, normalmente denominado *sumário ideal*. Essa forma de comparação tem sido a mais utilizada atualmente, pois pode ser mais facilmente automatizada. Uma vez tendo o sumário de referência, a avaliação da informatividade do sumário automático pode basear-se em várias medidas.

Um sumário de referência para um texto-fonte pode ser conseguido de várias formas:

- ele pode ser o sumário autêntico, isto é, o sumário produzido pelo próprio autor do texto-fonte;
- ele pode ser um sumário profissional, isto é, um sumário produzido a partir do texto-fonte por um escritor especialista em técnicas de sumarização;
- ele pode ser o extrato ideal, composto somente por sentenças mais representativas do texto-fonte.

É importante deixar claro que, apesar de usualmente se selecionar sentenças completas de um texto-fonte para formar seu sumário de referência, pode-se selecionar unidades com outros critérios, p.ex., trechos de sentenças, segmentos frasais, parágrafos, etc. A granularidade desejada depende do que se tem em mente ao avaliar um sistema de SA. Os sumários autêntico e profissional, segundo a definição acima, são os únicos resultantes da reescrita do texto-fonte (e, assim, ambos seriam *abstracts*, em nossa convença terminológica).

O extrato ideal é o melhor tipo de sumário de referência para a avaliação de sistemas de SA, pois, por conter somente sentenças do texto-fonte, pode ser comparado mais facilmente com um sumário automático, já que este também se origina do mesmo texto-fonte. No caso de extratos, a comparação com o extrato ideal pode ser automatizada; no caso de *abstracts*, podem ser necessárias etapas de revisão (humana) após o processamento.

O extrato ideal também pode ser produzido de várias formas:

- ele pode ser composto pelas sentenças do texto-fonte que mais se assemelhem às sentenças do sumário autêntico;
- ele pode ser composto pelas sentenças do texto-fonte que mais se assemelhem às sentenças do sumário profissional;
- ele pode ser composto pelas sentenças do texto-fonte julgadas por humanos como essenciais para compor um sumário do texto.

Na busca pelas sentenças do texto-fonte que mais se assemelham às sentenças do sumário autêntico ou profissional (*abstracts*), pode-se fazer uso de várias medidas. A mais utilizada, sugerida por Salton (1989), é a medida do co-seno, baseada puramente na co-ocorrência de palavras. Dessa forma, para cada sentença dos *abstracts* procura-se pela sentença do texto-fonte que tenha mais palavras em comum com aquela. Ao final, a justaposição das sentenças selecionadas do texto-fonte forma o extrato ideal. Outra medida, sugerida por Teufel e Moens

(2002), utiliza, além da co-ocorrência de palavras, a ordenação das palavras nas sentenças. É importante ressaltar, entretanto, que essas medidas não são perfeitas, visto que não se realiza nenhum tipo de análise semântica das sentenças. A simples verificação de co-ocorrência e ordenação de palavras não garante que duas sentenças tenham o mesmo conteúdo informativo, podendo, eventualmente, introduzir erros na produção do extrato ideal. É por isso que se aconselha, quando possível, uma revisão humana dos extratos produzidos.

Com relação à construção do sumário de referência a partir das sentenças julgadas por humanos como essenciais, deve-se ressaltar a questão da baixa concordância entre os julgamentos humanos. Como vários experimentos têm mostrado (por exemplo, Mitra et al., 1997; Rath et al., 1961), juízes humanos, em geral, concordam muito pouco sobre as sentenças que devem fazer parte de um sumário. O que se costuma fazer é selecionar somente aquelas sobre as quais há maior concordância dos juízes. Por outro lado, Marcu (1999) ressalta que, apesar de ser possível uma baixa concordância *geral* entre juízes, é possível que, pelo menos na escolha das sentenças mais importantes, a taxa de concordância seja maior. Assim, seria preciso distinguir a forma de avaliar a validade dos resultados considerando também possíveis variações de julgamento dos próprios juizes.

Tendo em mãos o sumário de referência para um texto-fonte, as formas possíveis de se verificar a informatividade de um sumário automático são:

- Cálculo automático da precisão e cobertura do sumário automático em relação ao sumário de referência. Aplicável, preferencialmente, a *extratos ideais* como sumários de referência e a sistema de SA extrativos. A precisão (P) e a cobertura (do inglês, *recall*) (R) são dadas pelas seguintes fórmulas:

$$P = \frac{número\ de\ sentenças\ do\ sumário\ automático\ presentes\ no\ sumário\ de\ referência}{número\ de\ sentenças\ do\ sumário\ automático}$$

$$R = \frac{número\ de\ sentenças\ do\ sumário\ automático\ presentes\ no\ sumário\ de\ referência}{número\ de\ sentenças\ do\ sumário\ de\ referência}$$

  A precisão indica quantas sentenças do sumário de referência o sumário automático possui em relação a todas as sentenças que ele contém; a cobertura indica quantas sentenças do sumário de referência o sumário automático possui em relação a todas as sentenças que ele deveria possuir. Uma outra medida, a *f-measure*, combina as medidas de precisão e cobertura, resultando em uma medida única de eficiência do sistema: quanto mais próxima essa medida for de 1, maior a capacidade do sistema em produzir sumários ideais. A fórmula da *f-measure* é a seguinte:

$$f - measure = \frac{2 \times P \times R}{P + R}$$

- Preferencialmente para um sistema extrativo, em vez de precisão e cobertura, pode-se utilizar a medida de *utilidade* de Radev et al. (2000). Para seu cálculo, pede-se a juízes humanos que dêem notas variando em uma determinada escala (de 1 a 9, por exemplo) para todas as sentenças do texto-fonte, que expressem sua importância para compor um sumário (uma nota é chamada de *ponto de utilidade*). Calcula-se, então, a nota geral do sumário de referência pela soma das notas de suas sentenças. Deste modo, a nota geral do sumário automático deve ser próxima (ou mesmo maior) do que a nota do sumário de referência para que ele seja considerado um sumário suficientemente informativo.

A vantagem dessa medida é que ela é mais flexível do que as medidas de precisão e cobertura, pois não penaliza tanto um sumário automático quando este não possui alguma(s) das sentenças do sumário de referência. Assim, o julgamento não se dá em relação à presença ou ausência das sentenças do sumário de referência no sumário automático, mas em relação à importância (numérica) das sentenças. É importante ressaltar, entretanto, que também pode haver uma baixa concordância entre os juízes humanos em sua atribuição de notas às sentenças do texto-fonte.

- Outra medida, mais genérica e, portanto, aplicável a outros métodos de SA além dos extrativos (i.e., àqueles que produzem *abstracts*), é a medida de *conteúdo* (Salton e McGill, 1983). Por ela, verifica-se a parcela do conteúdo do sumário de referência que é transmitida pelo sumário automático, não levando em consideração valores numéricos (quantidade de sentenças ou tamanho dos sumários, por exemplo). Essa verificação pode ser manual, subjetiva, ou auxiliada por processos automáticos. O cálculo da medida do co-seno é um exemplo de processo automático que subsidia a identificação de passagens com mesmo conteúdo.
  Pela medida de conteúdo pode-se verificar, também, quanto do conteúdo do próprio texto-fonte é preservado no sumário automático, como fizeram Brandow et al. (1994).

É importante notar que as medidas de precisão e cobertura e de utilidade também podem ser aplicadas para outros métodos além dos extrativos, necessitando, somente, de algum esforço humano para definir a forma de cômputo das medidas.

O problema em se usar sumários de referência para a avaliação da informatividade de sumários automáticos é que o sumário de referência pode ser inadequado ou até mesmo ruim. Os sumários autênticos, por exemplo, podem conter informação não apresentada no texto-fonte ou mesmo ser pouco informativos. Nesses casos, a comparação fica prejudicada, já que não há um mecanisms de compreensão para concluir por um fator comum entre variações desse tipo. É importante, portanto, selecionar as fontes utilizadas para a avaliação. Kupiec et al. (1995), por exemplo, retiraram dos sumários de referência as informações que não estavam nos textos-fonte correspondentes, resultando nos chamados sumários *gold-standard*, considerados os sumários ideais.

Além da comparação do sumário automático com o sumário de referência ou com o texto-fonte, há outras formas de se verificar a informatividade dos sumários automáticos. Mani (2001) sugere que, se um sumário for informativo, ele deve preservar os mesmos conceitos-chave de seu texto-fonte, os quais podem ser expressos por suas próprias palavras-chave. Assim, pode-se verificar, por exemplo, se as palavras-chave fornecidas pelo autor do texto-fonte ou aquelas calculadas por algum concordanceador[45] também são as palavras-chave do sumário automático ou se, pelo menos, estão presentes nele. Pardo e Rino (2002), em um outro tipo de avaliação, pedem a juízes humanos que dêem notas a sumários automáticos de acordo com a preservação da idéia principal dos textos-fonte correspondentes. As notas, nesse caso, indicam se o sumário preserva, preserva parcialmente ou mesmo não preserva a idéia principal. Essa proposta de avaliação se baseia na hipótese de que um sumário minimamente informativo deve transmitir, pelo menos, a idéia principal do texto-fonte.

Como se pode perceber, a subjetividade e a concordância dos julgamentos humanos é um grande desafio na avaliação de sistemas de SA, tanto na qualidade como na informatividade. O problema da subjetividade pode ser amenizado pela especificação clara de critérios de avaliação, pelo estabelecimento de escalas numéricas objetivas, em vez de conceitos abstratos, e pelo treinamento dos juízes, que, apesar de custoso, normalmente

[45] Um *concordanceador* é um programa que calcula dados estatísticos para um texto de entrada, por exemplo, a lista de palavras-chave, a freqüência de cada palavra do texto, etc.

produz bons resultados. Em relação à concordância entre os julgamentos, sejam quais forem os objetivos dos mesmos, a avaliação pode não ser válida ou estar comprometida se houver uma baixa concordância entre os juízes. É por esse motivo que alguns métodos para se medir a concordância entre julgamentos foram propostos. A medida Kappa (Siegel e Castellan, 1988) é a mais conhecida, bastante utilizada pelos trabalhos atuais na Lingüística Computacional.

15.4. Avaliação extrínseca

Como já mencionado, a avaliação extrínseca visa avaliar um sistema em uso, para a realização de alguma tarefa específica. Para a SA, a avaliação extrínseca pode envolver, por exemplo, os contextos de categorização de textos, recuperação de informação ou perguntas e respostas. Algumas dessas avaliações são discutidas nesta seção. É importante lembrar a avaliação extrínseca é uma forma de validação do sistema em uso: pode-se validar, por exemplo, sua metodologia e/ou seu modelo lingüístico-computacional.

15.4.1. Categorização de documentos

Em uma tarefa de categorização de documentos, muito realizada em sites de notícias e necessária para catalogação de documentos em bibliotecas, por exemplo, o objetivo é atribuir uma categoria/classe a um dado documento. Normalmene, essa atribuição é feita por juízes humanos. Em sites de notícias, por exemplo, devem-se classificar as notícias para enquadrá-las em suas devidas seções, como “ciência”, “economia”, “informática”, etc.

Em uma avaliação extrínseca de um sistema de SA para a tarefa de categorização de documentos, o que se costuma fazer é pedir aos juízes que categorizem os documentos lendo somente os sumários correspondentes. A seguir, verifica-se o tempo necessário para a realização da tarefa e a taxa de acerto dos juízes. Idealmente, espera-se que a taxa de acerto não degrade em relação à tarefa realizada de forma usual (isto é, lendo-se os textos-fonte em vez dos sumários) e que o tempo de realização da tarefa diminua. Em relação à SA, o objetivo deste tipo de avaliação é verificar se os sumários apresentam informação suficiente para a correta classificação dos textos-fonte, a partir de sua categorização. Esse tipo de avaliação foi proposto na conferência SUMMAC, como relatam Mani et al. (1998).

15.4.2. Recuperação de informação

Em uma tarefa de recuperação de informação, o objetivo é recuperar documentos que abordem um determinado tópico, como se costuma fazer em sites de busca de informação na web. O que se faz, neste caso, é pesquisar os documentos de uma base de dados em busca daqueles cujo tópico coincida com o tópico indicado pelo usuário. Similarmente à categorização de documentos, essa tarefa também é, muito freqüentemente, realizada por seres humanos, os quais desejam selecionar, entre vários documentos, aqueles que abordam algum assunto de seu interesse.

Na avaliação extrínseca, um sistema de SA é utilizado para gerar os sumários dos documentos da base que será pesquisada. A busca de documentos é feita, então, com base nos sumários dos documentos. Ela pode ser automática ou mesmo manual. O sucesso da busca é, então, avaliado por juízes humanos. Depois, verifica-se a taxa de acerto na recuperação e o tempo necessário para a busca, como no caso da categorização. Novamente, espera-se que a taxa de acerto se mantenha e que o tempo de busca diminua. Nesses casos, mede-se se os sumários realmente preservam todos os tópicos relevantes dos documentos para que a busca

possa ser feita. Além da conferência SUMMAC (Mani et al., 1998), vários outros trabalhos abordaram esta avaliação (por exemplo, Tombros e Sanderson, 1998; Jing et al., 1998; Brandow et al., 1994; etc.).

15.4.3. Perguntas e respostas

Em uma avaliação extrínseca de perguntas e respostas, tem-se por objetivo verificar se o sistema de SA produz sumários informativos ou não. Nessa avaliação, dada uma base de documentos, elaboram-se algumas perguntas de múltipla escolha para cada texto. A seguir, aplica-se o sistema de SA aos documentos para produzir os sumários correspondentes. Por fim, procede-se então à avaliação propriamente dita. Primeiro, pede-se a juízes humanos que respondam às questões sem ler os textos-fonte nem os sumários; a seguir, pede-se aos juízes que leiam os sumários e respondam as questões; em um último passo, pede-se aos juízes que leiam os textos-fonte e respondam as mesmas perguntas.

A hipótese principal, neste caso, é que, se os sumários forem devidamente informativos, os juízes conseguirão responder as perguntas satisfatoriamente lendo somente os sumários. Costuma-se pedir aos juízes que repitam o procedimento lendo os textos-fonte e não lendo nada pelas seguintes razões: se, sem ler nada, os juízes conseguem responder corretamente algumas perguntas, isso indica que as mesmas são de senso comum e, portanto, devem ser excluídas da avaliação; se, mesmo lendo o texto-fonte completo, os juízes não conseguem responder algumas perguntas, isso indica que, muito provavelmente, os sumários também não irão respondê-las satisfatoriamente. Neste caso, elas não servem para avaliá-los, novamente, e, assim, devem ser excluídas da avaliação. Ao final, restarão as perguntas e respostas que realmente servirão para avaliar a informatividade dos sumários. Dentre os trabalhos que aplicaram esta avaliação, destacam-se os trabalhos de Morris et al. (1992) e Hovy e Lin (2000).

Em geral, as avaliações extrínsecas, assim como as intrínsecas, também apresentam diversos desafios para sua realização, por exemplo:

- as avaliações extrínsecas normalmente são on-line, isto é, precisam de juízes humanos, e, como já discutido, a avaliação on-line é custosa;
- por normalmente serem on-line, as avaliações extrínsecas pedem por documentos relativamente curtos para facilitar o trabalho dos juízes humanos. Entretanto, se forem muito curtos, sequer há necessidade de sumários;
- as avaliações extrínsecas, diferentemente das intrínsecas, não indicam pontos específicos em que os sistemas de SA utilizados podem ser aprimorados. Isso ocorre porque elas medem o desempenho das tarefas nas quais os sistemas de SA estão inseridos, e não os sistemas propriamente ditos;
- às vezes, é difícil criar tarefas extrínsecas que modelem adequadamente situações do mundo real e, ao mesmo tempo, sejam passíveis de medição e possíveis de serem realizadas por juízes humanos.

15.5. Estudo de caso: avaliação do GistSumm

Como estudo de caso, será apresentada a avaliação do GistSumm – ***GIST SUMM****arizer*. Será mostrado o raciocínio por trás da definição da forma de avaliação e as conclusões inferidas com base nos resultados da avaliação.

Como já explicado anteriormente, as hipóteses relativas ao processo de sumarização do GistSumm são: (I) é possível determinar a *gist sentence* de um texto por meio de métodos estatísticos simples ou, pelo menos, se aproximar dela e (II) com base na *gist sentence*, é possível construir bons extratos. Portanto, a avaliação do GistSumm deve buscar a validação dessas hipóteses.

Para avaliar a hipótese I, deve-se especificar algum procedimento em que seja possível mostrar se a *gist sentence* pode ou não ser determinada pelos métodos estatísticos simples do GistSumm. Como relatam Pardo et al. (2003), a idéia foi, então, pedir a juízes humanos que determinassem as *gist sentences* de alguns textos (10, no total) e, então, verificar se o GistSumm identificava ou não estas *gist sentences* e, caso elas não fossem identificadas como tais, se elas eram incluídas ou não nos extratos produzidos automaticamente. Dada a subjetividade do julgamento humano, a *gist sentence* escolhida para cada texto foi aquela que recebeu mais votos dos juízes. A taxa de compressão utilizada foi de 60%.

É importante ressaltar que utilizar mais de 10 textos para essa avaliação a tornaria muito custosa, pois, quanto mais textos, mais tempo os juízes levariam para ler e identificar as *gist sentences* dos textos. Sempre que se utilizam juízes humanos, é importante balancear a quantidade de esforço necessário para a realização da tarefa, a capacidade dos juízes e o tempo disponível para a realização da tarefa.

Os resultados obtidos com a avaliação acima foram os seguintes:

- Utilizando o método das palavras-chave: as *gist sentences* escolhidas pelos juízes foram identificadas em 20% dos casos (ou seja, 2 textos); em 50% dos casos (5 textos), o gistsumm escolheu *gist sentences* muito próximas das *gist sentences* indicadas pelos juízes; para os 30% restantes (3 textos), o gistsumm não conseguiu sequer uma aproximação das *gist sentences* indicadas pelos juízes, ou seja, falhou. No total, as *gist sentences* indicadas pelos juízes foram incluídas nos extratos em 70% dos casos (7 textos).
- Utilizando o método TF-ISF: as *gist sentences* escolhidas pelos juízes foram identificadas em 20% dos casos (ou seja, 2 textos); em 10% dos casos (1 texto), o gistsumm escolheu *gist sentences* vagamente próximas das *gist sentences* indicadas pelos juízes; para os 70% restantes (7 textos), o GistSumm não conseguiu sequer uma aproximação das *gist sentences* indicadas pelos juízes, ou seja, falhou. No total, as *gist sentences* indicadas pelos juízes foram incluídas nos extratos em 30% dos casos (3 textos).

Como resultado, tem-se que o método das palavras-chave é satisfatório para a determinação das *gist sentences*, validando a hipótese I, enquanto o método TF-ISF não. Por esse motivo, pode-se descartar o método TF-ISF e futuros investimentos nele, pois, se ele sequer consegue identificar as *gist sentences*, ele não gerará bons extratos.

Essa avaliação pode ser classificada como **intrínseca**, pois analisa a qualidade do sistema em si, **glass-box**, pelo fato de analisar um dos componentes do GistSumm (o "módulo" que determina a *gist sentence*), **comparativa**, por se comparar dois métodos de determinação de *gist sentences*, e **off-line**, pelo fato de não utilizar o julgamento humano na avaliação. É importante ressaltar que, apesar dos juízes terem sido utilizados para detectar as *gist sentences* dos textos, eles não foram utilizados para julgar os extratos (etapa esta que, de fato, caracteriza a avaliação como off-line ou não).

Para avaliar a hipótese II, se os extratos produzidos pelo GistSumm são bons ou não (usando somente o método das palavras-chave para determinar as *gist sentences*), Pardo et al. tiveram que recorrer ao julgamento humano. Eles estabeleceram uma escala de pontuação para medir duas características dos extratos, a preservação da idéia principal dos textos-fonte e a textualidade (vide Quadro 1). Dessa forma, por exemplo, caso um juiz achasse que um extrato preservou a idéia principal do texto-fonte, mas não apresentou textualidade, então ele deveria dar nota 7 ao extrato. A idéia principal tem a ver com a informatividade dos extratos,

ou seja, para que estes sejam minimamente informativos, eles devem preservar a idéia principal dos textos-fonte, pelo menos. A textualidade, por sua vez, engloba também a qualidade dos extratos, pois verifica a coesão deste, sua "fluência" durante a leitura.

Os resultados obtidos foram: 55% dos extratos gerados pelo GistSumm estavam acima da média e 14% dos extratos estavam na média; 50% dos extratos preservaram totalmente a idéia principal e 40% preservaram parcialmente; 50% dos extratos apresentaram textualidade total e 35% apresentaram textualidade parcial. Dessa forma, 90% dos extratos preservaram totalmente ou parcialmente a idéia principal e 85% dos extratos apresentaram textualidade total ou parcial. Portanto, pode-se considerar que a hipótese II foi validada. Essa avaliação pode ser classificada como **intrínseca**, **black-box**, **autônoma** e **on-line**.

Quadro 1. Escala de pontuação dos extratos

Idéia principal	Textualidade	Nota
Preservada	Ok	9
Preservada	±	8
Preservada	Sem	7
Parcialmente preservada	Ok	6
Parcialmente preservada	±	5
Parcialmente preservada	Sem	4
Não preservada	Ok	3
Não preservada	±	2
Não preservada	Sem	1

Em uma última avaliação, o GistSumm participou da conferência DUC realizada no início de 2003, já que, por ser uma conferência de avaliação de caráter internacional, ela é confiável e, portanto, dá mais validade aos resultados obtidos com a avaliação do sistema. Na primeira etapa da avaliação, de natureza **extrínseca**, verificou-se se os extratos produzidos pelo GistSumm eram "úteis" ou não para a tarefa em que um usuário tem que selecionar que documentos ler com base nos seus sumários. Dados 624 textos-fonte, para cada extrato produzido pelo GistSumm (com uma média de 38 palavras), juízes humanos, especialistas em técnicas de recuperação de informação, deram notas de 0 a 4 aos extratos, onde 0 indicava um extrato inútil e 4 um extrato perfeito que poderia até mesmo substituir o texto-fonte. Nesta avaliação, o GistSumm conseguiu uma nota média de 3.12, o que caracteriza seus extratos como sendo muito bons. Em uma outra etapa, agora **intrínseca**, para verificar a informatividade dos extratos, juízes humanos calcularam a cobertura (*recall*) dos extratos em relação a sumários profissionais. O GistSumm atingiu uma cobertura média de 51%. Ambas as avaliações anteriores são consideradas **black-box**, **autônomas** e **on-line**. Vale citar que, em uma última etapa da DUC, foi realizada uma avaliação **comparativa** entre os sistemas de SA participantes da conferência, porém, o GistSumm não participou desta etapa.

15.6. Considerações finais sobre avaliação de sistemas de SA

A avaliação de sistemas de SA é um assunto muito amplo. Há diversas formas de se avaliar um sistema de SA, podendo-se focalizar suas características computacionais, o design de sua interface e, mais importante, os resultados produzidos, que, neste caso, são os próprios sumários.

Avaliar, em geral, é um processo custoso, ainda mais pelo fato de precisar, com freqüência, do julgamento humano. Métodos automáticos de avaliação existem, como discutido nesta seção, mas apresentam diversos problemas e não são tão satisfatórios como o

julgamento humano. Mesmo o julgamento humano pode ser problemático, dada a subjetividade desta tarefa e a baixa concordância entre juízes. Apesar das dificuldades, padrões e métricas de avaliação de sistemas de SA têm surgido na literatura, assim como as conferências internacionais de avaliação têm se tornado cada vez mais importante.

Mani (2001) afirma que a avaliação de sistemas de SA tem que nortear o desenvolvimento de tecnologia na área e ser norteada por esse mesmo processo. Entretanto, avaliar não é simplesmente seguir um "livro de receitas", pois depende das necessidades e características de cada sistema de SA e dos objetivos dos desenvolvedores do sistema, que pode ser desde melhorar o "estado da arte", em termos de modelos e métodos de sumarização, até adequar os sistemas a tarefas específicas do mundo real.

A avaliação em SA é um tema desafiador e necessário para o desenvolvimento da pesquisa, que ainda precisa ser bastante trabalhado em busca de padrões e metodologias adequadas. Citando o próprio Mani (2001, p. 224): "*if all we do is evaluation, evaluation is all we will do!*"

Referências da Parte III

Aretoulaki, M. (1996). *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis of Texts For Their Automatic Summarisation*. PhD. Thesis. University of Manchester.

Barzilay, R.; Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In the *Proc. of the Intelligent Scalable Text Summarization Workshop*, Madri, Spain. Also In I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization*. MIT Press, pp. 111-121.

Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-365.

Black, W.J.; Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology.

Boguraev, B.; Kennedy, C. (1997). Salience-Based Content Characterisation of Text Documents. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop*, pp. 2-9. ACL/EACL'97 Joint Conference. Madrid, Spain.

Borko, H.; Bernier, C.L. (1975). *Abstracting Concepts and Methods*. Academic Press. San Diego, CA.

Braga, A.P.; Ludermir, T.B.; Carvalho, A.C.P.L.F. (2000). *Redes Neurais Artificiais: Teoria e aplicações*. LTC - Livros Técnicos e Científicos Editora S.A, Rio de Janeiro.

Brandow, R.; Karl, M.; Rau, L.F. (1994). Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing & Management*, Vol. 31, N. 5, pp. 675-685.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Corston-Oliver, S. (1998). *Computing Representations of the Structure of Written Discourse*. PhD Thesis, University of California, Santa Barbara, CA, USA.

Cremmins, E.T. (1996). *The Art of Abstracting*. Information Resource Press. Arlington, Virginia.

Edmundson, H.P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, 16, pp. 264-285.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221-233.

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

Halliday, M.A.K.; Hasan, R. (1976). *Cohesion in English*. Longman.

Hoey, M. (1983). *On the Surface of Discourse*. George Allen & Unwin Ltd.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

Hovy, E.; Lin, C-Y. (1997). "Automated Text Summarization in SUMMARIST", Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference. Madrid, Spain, p. 18-24.

Hovy, E.H.; C-Y. Lin. (2000). Automated Text Summarization and the SUMMARIST System. In the *Proceedings of the TIPSTER Text Program, Phase III*, pp. 197-214.

Jing, H.; Barzilay, R. McKeown, K.; Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In the *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*.

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

Jordan, M. P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W. C. Mann and S. A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Kupiec, J.; Petersen, J.; Chen, F. (1995). A trainable document summarizer. In Edward Fox, Peter Ingwersen, & Raya Fidel (eds.), *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval*, pp. 68-73, Seattle, WA, EUA. July.

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, Atibaia, SP.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

Mani, I. (2001). Summarization Evaluation: An Overview. In the *Proceedings of the Workshop on Automatic Summarization*. Pittsburgh, Pennsylvania.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mani, I.; Firmin, T.; House, D.; Chrzanowski, M.; Klein, G.; Hirschman, L.; Sundheim, B.; Obrst, L. (1998). *The TIPSTER Text Summarization Evaluation*. Final Report.

Mani, I.; Maybury, M.T. (1999), eds. *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

Mann, W.C.; Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8 (3), pp. 243-281.

Marcu, D. (1997b). The Rhetorical Parsing of Natural Language Texts. In the *Proc. of the ACL/EACL'97 Joint Conference*, pp. 96-103. Madrid, Spain.

Marcu, D. (1997a). From Discourse Structures to Text Summaries. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop*, pp. 82-88. ACL/EACL'97 Joint Conference. Madrid, Spain.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization,* pp. 123-136. The MIT Press.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Martins, C.B. (2002). *UNLSumm: Um Sumarizador Automático de Textos UNL*. Dissertação de Mestrado, DC/UFSCar, São Carlos.

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.

Miike, S.; Itoh, E.; Ono, K.; Sumita, K. (1994). A full text-retrieval system with a dynamic abstract generation function. In W. Bruce Croft and C.J. van Rijsbergen (eds.), *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval*, pp. 152-161, July.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communication of the Association for Computing Machinery* 38 (11), pp. 39-41.

Minel, J.L.; Nugier, S.; Piat, G. (1997). How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN. In I. Mani and M. Maybury (eds.), *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*.

Mitchell, T.M. (1997). *Machine Learning*. McGraw Hill, New York.

Mitra, M.; Singhal, A.; Buckley, C. (1997). Automatic text summarization by paragraph extraction. In the *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*.

Moore, J.D.; Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol. 19, No. 4, pp. 651-694.

Morris, J.; Hirst, G. (1991). Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics,* 17(1): 21-48.

Morris A.; Kasper, G.; Adams, D. (1992). The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research*, Vol. 3, N. 1, pp. 17-35.

Nunes, M.G.V.; Vieira, F.M.V; Zavaglia, C.; Sossolete, C.R.C.; Hernandez, J. (1996). *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Série de Relatórios Técnicos do ICMC-USP, no. 42.

Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).

Paice, C.D.; Jones, P.A. (1993). The identification of important concepts in highly structure technical papers. In R. Korfaghe, E. Rasmussen, and P. Willett (eds.), Proc. of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 69-78. ACM Press, June.

Pardo, T.A.S. (2002a). *GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos*. Série de Relatórios do NILC. NILC-TR-02-13.

Pardo, T.A.S. (2002b). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Pardo, T.A.S. (2002c). *Descrição do DMSumm: um Sumarizador Automático Baseado em um Modelo Discursivo*. Série de Relatórios do NILC (DC-UFSCar). NILC-TR-02-02.

Pardo, T.A.S.; Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

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

Pinheiro, G.M. e Aluísio, S.M. (2003). *Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.

Pollock, J.J.; Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Compute Sciences* 15(4): 226-232.

Pollock, J.J.; Zamora, A. (1999). Reedição de (Pollock e Zamora, 1975).

Radev, D.R.; Jing, H.; Budzikowska, M. (2000). Summarization of multiple documents: clustering, sentence extraction, and evaluation. In the *Proceedings of the Workshop on Automatic Summarization*, pp. 21-30. Seattle, WA.

Rath, G.J.; Resnick, A.; Savvage, R. (1961). The formation of abstracts by the selection of sentences. *American Documentation*, Vol. 12, N. 2, pp. 139-141.

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

Rino, L.H.M.; Scott, D. (1994). *Automatic generation of draft summaries: heuristics for content selection.* ITRI Techn. Report ITRI-94-8. University of Brighton, England.

Rowley, J. (1982). *Abstracting and Indexing*. Clive Bingley, London.

Saggion, H.; Lapalme, G. (2000). Concept identification and presentation in the context of technical text summarization. In the *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, pp. 1-10. Seattle, WA.

Salton, G. (1988). *Automatic Text Processing*. Reading, MA: Addison-Wesley.

Salton, G. (1989) *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

Salton, G.; McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill. New York.

Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. (1997). Automatic Text Structuring and Summarization. *Information Processing & Management*, 33(2), pp. 193-207.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, Vol. 8. Cambridge University Press.

Sparck Jones, K. (1993a). *Discourse Modelling for Automatic Summarisation*. Tech. Report No. 290. University of Cambridge. UK, February.

Sparck Jones, K. (1993b). What might be in a summary? In Krause Knorz and Womser-Hacker (eds.), *Information Retrieval 93*, pp. 9-26. Universitatsverlag Konstanz. June.

Sparck Jones, K.; Galliers, J.R. (1996). Evaluating Natural Language Processing Systems. *Lecture Notes in Artificial Intelligence*, Vol. 1083.

Sparck Jones, K. (1997). "Summarising: Where are we now? Where should we go?" *Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference*. Madrid, Spain, p. 1.

Siegel, S.; Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Teufel, S.; Moens, M. (1999). Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In Inderjeet Mani and Mark T. Maybury (Eds.), *Advances in Automatic Text Summarization*. Massachusetts Institute of Technology Press.

Teufel, S.; Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, Vol. 28, N. 4, pp. 409-445.

Tombros, A.; Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In the Proceedings of the 21st ACM SIGIR Conference, pp. 2-10.

Uchida, H. (1997). *DeConverter Specification, Version 1.0*. Tech. Rep. UNL-TR1997-010, UNU/IAS/UNL Center, Tokyo, Japan.

Uchida, H. (2000). *Universal Networking Language: An Electronic Language for Communication, Understanding and Collaboration*. UNL Center, IAS/UNU, Tokyo (também disponível no site www.unl.ias.unu.edu).

White, J. S.; Doyon, J. B.; Talbott, S. W. (2000). Task Tolerance of MT Output in Integrated Text Processes. In *ANLP/NAACL 2000: Embedded Machine Translation Systems*, pp. 9-16. Seattle, WA

Winter, E.O. (1976). *Fundamentals of Information Structure*. Hatfield Polytechnic, Hertfodshire, England.

Winter, E.O. (1977). A Clause-Relational Approach to English Texts. A Study of Some Predictive Lexical Items in Written Discourse. *Structional Science*, Vol. 6, N. 1, pp. 1-92.

Winter, E.O. (1979). Replacement as a Fundamental Function of the Sentence in Context. In *Forum Linguitics*, Vol. 4, N. 2, pp. 95-133.

Witten, I.H.; Moffat, A.; Bell, T.C. (1994). Managing Gigabytes. *Van Nostrand Reinhold*. New York.