Lingüística computacional: princípios e aplicações

Renata Vieira[1] , Vera Lúcia Strube de Lima[2]

[1] Centro de Ciências da Comunicação, Centro de Ciências Exatas e Tecnológicas
UNISINOS Av. Unisinos, 950 CEP 93022-000 São Leopoldo RS

[2] Faculdade de Informática
PUCRS Av. Ipiranga, 6681 CEP 90619-900 Porto Alegre RS

renata@exatas.unisinos.br vera@inf.pucrs.br

Resumo *A lingüística computacional é a área de conhecimento que explora as relações entre lingüística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural. Neste curso são caracterizados os conhecimentos relativos à língua utilizados na construção de tais sistemas, bem como, as técnicas empregadas para o processamento dos diferentes níveis lingüísticos (lexical, sintático e semântico-pragmático). Uma discussão sobre desenvolvimento da área e a multiplicidade de aplicações e produtos decorrentes das pesquisas em lingüística computacional é apresentada.*

Abstract *Computational linguistics is an area of research that is based on the connection between linguistics and computer science. This union enables the development of systems which are capable of interpreting and producing information that is presented in natural language. In this tutorial we review both the linguistic knowledge that is used for the construction of such systems and the computing techniques applied to various levels of language processing (lexical, syntactic, semantic and pragmatical). A discussion about the research in the area and the great number of applications and products resulting from it is presented.*

1. Introdução

O desenvolvimento da informática proporcionou, nas últimas décadas, grandes mudanças nos estudos das ciências em geral. A computação, no caso particular do estudo das línguas naturais, possibilitou o surgimento de novas abordagens a problemas descritivos e práticos das línguas que antes não podiam ser tratados adequadamente.

Uma destas abordagens é a lingüística baseada em corpus, que utiliza computadores para o armazenamento e acesso a textos escritos ou falados. Um corpus lingüístico legível por máquina pode ser rapidamente pesquisado para obtenção de informações a respeito da regularidade da língua, tais como freqüência de palavras, de formas ou de construções. Desta maneira pode-se obter dados a respeito da linguagem[1] real, em uso por falantes da língua, permitindo fazer comparações entre língua escrita e

[1] Os termos língua e linguagem são utilizados alternadamente ao longo desse trabalho sem uma distinção específica.

falada, entre os usos da língua em diferentes épocas, ou ainda, entre o português do Brasil e de Portugal, para citar alguns exemplos.

Outros trabalhos em lingüística computacional são voltados ao processamento da linguagem natural, isto é, à construção de programas capazes de interpretar e/ou gerar informação fornecida em linguagem natural. Para o processamento da língua natural, vários subsistemas são necessários para dar conta dos diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso. Alguns exemplos são apresentados a seguir.

Para ter uma comunicação efetiva, os usuários da língua costumam seguir certas convenções. Uma destas convenções permite ao falante nativo reconhecer uma seqüência de expressões como sendo uma sentença válida da língua. O processamento lingüístico, a esse nível, é tarefa dos analisadores sintáticos. Para verificar a validade de seqüências de palavras numa certa língua, o sistema precisa que a língua seja especificada por um léxico e uma gramática. O procedimento é similar à verificação de sintaxe de um programa em uma linguagem de programação, a sintaxe da língua natural é, no entanto, bem mais complexa e é preciso levar em consideração problemas particulares como o da concordância, por exemplo. Esse tipo de tratamento é útil ao desenvolvimento de corretores ortográficos e gramaticais. As aplicações desenvolvidas para lidar com a língua, porém, vão além do processamento sintático, como será mostrado a seguir.

Podemos, inicialmente, observar a diferença entre os sistemas que lidam com a língua escrita e a língua falada. Para lidar com a língua falada é necessária uma tecnologia especial que faz a interpretação da fala através da manipulação da representação de conhecimento fonético-fonológico.

Um outro aspecto da língua, diz respeito ao significado que é evocado por uma sentença válida. Uma sentença pode expressar o conhecimento de mundo ou uma intenção do falante em relação ao ouvinte. Para desenvolver sistemas com essas características é preciso recorrer a técnicas de representação do conhecimento e, em certas situações, especificar algoritmos capazes de estabelecer relações entre os diversos componentes e segmentos de um texto ou discurso. Esses são os sistemas de tratamento semântico da língua, que podem envolver a construção de um modelo de representação do domínio, correspondente à interpretação de um texto, ou podem lidar com questões mais pontuais, como reconhecer um sentido específico, dentro de um contexto, para palavras ambíguas (por exemplo, *banco* como instituição financeira ou *banco* como um artefato utilizado para sentar).

O significado da língua natural está sempre relacionado à situação de uso; no entanto, muitos modelos, utilizados para explicar e descrever o significado, procuram isolar esses fatores. A semântica, portanto, caracterizou-se como uma área de estudo que considera o significado das expressões lingüísticas de maneira independente de quem as usa ou de como são usadas. O estudo de questões relacionadas ao uso da língua acaba caracterizando uma outra área de conhecimento denominada pragmática.

Na pragmática são estudadas questões ligadas ao uso da linguagem, abordando-se aquilo que é relativo a quem usa e ao contexto de uso (a teoria dos atos de fala é um exemplo de tais estudos). Sistemas que trabalham nesse nível de representação costumam considerar o contexto lingüístico (discurso) na interpretação das expressões

da língua. O contexto lingüístico é o mais fácil de tratar computacionalmente, pois refere-se ao que é explicitado no texto. Sistemas que podem ser citados como exemplos são os de resolução de anáfora intersentencial e resolução de co-referência textual em geral. É mais difícil tratar computacionalmente o contexto imediato, ou contexto situacional de uma expressão, devido à dificuldade de se chegar a uma representação adequada do conhecimento compartilhado entre os participantes de uma conversação ou comunicação. Podemos considerar como conhecimento compartilhado, por exemplo, o conhecimento comum entre o leitor e o escritor de artigos de um jornal, que decorre de serem habitantes de uma mesma cidade.

Outros exemplos de aplicação de propósito mais geral, e que podem englobar no mesmo sistema vários dos níveis mencionados, são os sistemas de tradução automática, geração de resumos e extração de informação.

A área de lingüística computacional será aqui apresentada através de seus princípios gerais, de acordo com os diferentes níveis de conhecimento lingüístico. O resultado prático do desenvolvimento de pesquisas será ilustrado através da apresentação das suas principais aplicações. Na seção 2, caracterizaremos os níveis do estudo lingüístico, relacionados aos sons, formação das palavras e das frases, e o significado dos símbolos da linguagem. Na seção 3, uma introdução ao processamento da linguagem natural será apresentada, mostrando algumas formas de tratamento computacional para cada um dos níveis lingüísticos. Na seção 4, será apresentado um conjunto de aplicações da lingüística computacional. Na seção 5, apresentaremos uma introdução à lingüística computacional baseada em corpus, finalizando-se com a seção 6, a conclusão.

2. Áreas de estudos lingüísticos

Nesta seção serão apresentadas questões relativas aos diferentes níveis de estudo da linguagem: fonologia, morfologia, sintaxe, semântica e pragmática.

2.1. Fonética e fonologia

Fonética e fonologia são as áreas de estudo relacionadas ao sistema de sons de uma língua. A fonética está relacionada ao estudo da produção da fala humana, considerando as questões fisiológicas envolvidas, tais como a estrutura do aparelho fonador: mandíbula, laringe, boca, dentes e língua. Essa é uma estrutura bastante complexa, mais de 100 músculos estão envolvidos no controle direto e contínuo da produção das ondas sonoras da fala. Esse é o campo de estudo conhecido como fonética articulatória. Quando o estudo é mais voltado para as propriedades físicas das ondas sonoras da fala, entramos no campo da fonética acústica.

A fonologia é o estudo das regras abstratas e princípios envolvidos na organização, estrutura e distribuição dos sistemas de sons de uma determinada língua. Para se falar sobre os sons da língua é necessário um conjunto de símbolos que representem esses sons, pois a ortografia convencional apresenta problemas do tipo: diferentes sons são associados a uma mesma grafia e, por outro lado, diferentes grafias podem representar um mesmo som.

O domínio desse conhecimento é necessário ao desenvolvimento dos sistemas de reconhecimento e síntese de fala. O reconhecimento de fala envolve a interpretação de

ondas sonoras e a associação destas com elementos de fala, podendo reconhecer somente palavras isoladas dentro de um léxico pré-determinado (por exemplo, reconhecimento de números) ou reconhecer fala contínua de uma determinada língua (envolvendo o reconhecimento mais completo do léxico de uma língua e a delimitação/diferenciação entre várias palavras). A síntese da fala envolve a geração de uma saída sonora, a partir de um texto escrito de entrada. Uma das maiores dificuldades no desenvolvimento desse tipo de sistema é produzir pronúncia adequada e convincente, com sonoridade similar à fala humana.

2.2. Morfologia e sintaxe

A morfologia e a sintaxe estudam a constituição das palavras e dos grupos de palavras que formam os elementos de expressão de uma língua. A morfologia trata especificamente do conhecimento sobre a estrutura das palavras. Algumas palavras, como *árvore*, não podem ser quebradas em unidades menores, mas isso pode ocorrer com palavras como *árvores* ou *arvorezinhas*, por exemplo. Ou ainda palavras como *impossível*, ou *sobremesa*. As unidades constituintes das palavras são denominadas morfemas, e tais constituintes podem ser independentes, como em *árvore* ou dependentes como no caso dos sufixos (*s* em *árvores*) e prefixos (*im* em *impossível*).

Além de estudar a estrutura das palavras, em morfologia estuda-se a classificação das palavras em diferentes categorias, ou, conforme o termo popularmente conhecido na área, as palavras são classificadas em partes do discurso (*part-of-speech*, ou *POS*). Entre tais categorias encontramos os substantivos (*cachorro*), verbos (*correr*), adjetivos (*grande*), preposições (*em*), e advérbios (*rapidamente*). As palavras de uma mesma categoria compartilham várias propriedades em comum como, por exemplo, o tipo de plural (+ *s*) ou o tipo de diminutivo (+ *inho*). Os verbos e suas conjugações podem apresentar modificações regulares em vários casos. Na língua inglesa, os adjetivos podem ser acompanhados dos sufixos *er* e *est*, como em *big*, *bigger*, *biggest*, significando uma troca de adjetivo comum para um adjetivo comparativo ou superlativo. As categorias de palavras podem ainda ser divididas em classes abertas ou fechadas. As classes abertas são compostas por categorias que abrangem um grande número de palavras e podem, ainda, abrigar o surgimento de novas palavras. Classes dessas naturezas são os substantivos, verbos e adjetivos. As classes fechadas são aquelas que têm funções gramaticais bem definidas, tais como artigos, demonstrativos, quantificadores, conjunções e preposições.

Outra característica compartilhada entre as palavras de uma mesma categoria é a contribuição da palavra para o significado da frase que a contém. Por exemplo, substantivos podem ser usados para identificar um objeto ou conceito determinado, e adjetivos são usados para qualificar esse objeto ou conceito. Ainda a categoria pode dizer algo sobre a posição que as palavras podem ocupar nas frases. As palavras de determinada categoria podem ser usadas como base de um determinado grupo (ou sintagma). Tais palavras são chamadas de núcleo e identificam o tipo de objeto ou conceito que o sintagma descreve. Por exemplo, os sintagmas nominais possuem por núcleo um substantivo (ou nome); em *o cachorro, o cachorro raivoso* ou em *o cachorro raivoso do canil*, temos sintagmas nominais que descrevem o mesmo tipo de objeto. Da mesma forma, os sintagmas adjetivais *faminto, muito faminto, faminto como um cavalo*, descrevem um mesmo tipo de qualidade.

O reconhecimento das categorias das palavras é um problema básico em lingüística computacional. Muitas aplicações são desenvolvidas com base nessa informação inicial. Para se fazer a análise da estrutura das sentenças, por exemplo, é necessário que primeiramente se faça o reconhecimento das categorias. Os sistemas que realizam este tipo de tarefa são denominados etiquetadores de categorias gramaticais (ou *POS taggers*): dado um texto, esse texto é devolvido com o acréscimo, a cada palavra, de uma etiqueta com informação a respeito de sua categoria gramatical.

Uma vez reconhecida a categoria de uma palavra, o próximo passo na análise da linguagem natural, é verificar se a estrutura das frases é válida e reconhecer, dentro dessa estrutura, os constituintes da frase. Assim como palavras de uma mesma categoria, as estruturas das frases também compartilham determinadas propriedades, e é por essa razão que os falantes da língua podem reconhecer e produzir sentenças que nunca foram ouvidas antes. Esse conhecimento lingüístico referente à organização das palavras de uma frase em uma determinada ordem pode ser caracterizado por uma gramática, consistindo de um conjunto finito de regras e princípios. Essa ordem identifica a composição de constituintes que têm funções bem definidas na frase, como, por exemplo, sujeito e predicado. Correspondem a essas funções agrupamentos de palavras que obedecem a uma mesma regra de formação. Por exemplo, o sujeito é geralmente identificado por um grupo de palavras que constituem um sintagma nominal; o predicado é geralmente dado através de um sintagma verbal, que por sua vez é constituído de verbo e objeto, sendo que esse objeto é representado por outro sintagma nominal ou preposicional.

Através do reconhecimento da estrutura da frase é possível identificar quais expressões dizem respeito ao sujeito da frase, qual relação ou ação está sendo afirmada (ou seja, qual é o predicado da frase) e, para o predicado, identificar os objetos e outros complementos indicando, por exemplo, modo ou tempo da ação/relação. Além disso, é através da análise sintática que se pode verificar se a concordância estabelecida pelas regras da língua está sendo obedecida pela frase. Outra questão relacionada à estrutura da frase é a interpretação: diferentes possibilidades de combinações entre os constituintes de uma mesma frase podem ter diferentes interpretações (fenômeno denominado ambigüidade). Nos exemplos a seguir, podemos verificar a possibilidade de diferentes interpretações para as frases:

O homem viu o menino com o telescópio.

Ele entrou na sala de muletas.

As diferentes interpretações (*o menino com o telescópio* ou *viu com o telescópio*; *a sala de muletas* ou *entrou de muletas*) não são devidas à presença de ambigüidade nas palavras mas sim na estrutura. Diagramas em forma de árvore costumam ser usados para representar a constituição das frases de acordo com as regras de formação estabelecidas pela gramática.

Como podemos ver, apesar de estarem separados em diferentes tipos de problemas com diferentes abordagens e tratamentos, existe uma forte ligação entre os subsistemas da língua: para fazer a análise sintática requer-se informações morfológicas, e o resultado da análise sintática trará conseqüências para a interpretação de uma frase (conseqüências estas já no campo da semântica, apresentado na seção seguinte).

Sistemas que realizam a análise estrutural das frases e seus constituintes são os analisadores sintáticos (comumente conhecidos por sua denominação em inglês, *parsers*). Esses sistemas reconhecem estruturas válidas a partir de um léxico que define o vocabulário da língua e um conjunto de regras que definem a gramática da língua. Na área de lingüística computacional, muitos trabalhos são voltados ao problema da análise sintática. Um problema que ainda não está completamente solucionado.

2.3. Semântica e pragmática

Reconhecer se uma determinada seqüência de palavras está de acordo com as regras e princípios de formação de frases e sintagmas da língua é uma das ações envolvidas nos processos de interpretação e geração da linguagem natural. Associado a um enunciado bem formado está o seu significado, que pode ser uma proposição sobre os fatos do mundo ou, ainda, pode expressar o propósito ou a intenção do falante. A semântica tem como objeto de estudo o significado das expressões da linguagem natural e a pragmática irá estudar as relações dos significados com o contexto da enunciação.

A semântica aborda o significado das expressões de maneira mais independente de quem as usa ou de como são usadas essas expressões. O estudo do significado pode ser centralizado no significado das palavras, através da semântica lexical, ou no valor verdade de uma proposição, através da semântica lógica.

A semântica lexical considera as propriedades referentes a cada uma das unidades, ou seja, as palavras de uma língua, no léxico. Um dos primeiros problemas a serem considerados é o fato de algumas palavras apresentarem múltiplos sentidos. O verbo *ir*, por exemplo, apresenta 37 diferentes definições, no Dicionário Aurélio Básico da Língua Portuguesa. Por outro lado, como se pode observar em uma leitura do dicionário, essas definições são dadas em termos de outros verbos (por exemplo, *ir* = *partir*) e desse modo temos dois verbos compartilhando o mesmo sentido.

Para lidar com os sentidos, é comum organizá-los em classes de objetos, de acordo como usualmente classificamos as coisas do mundo. Tais classificações, taxonomias ou ontologias, têm sido de interesse desde o tempo de Aristóteles (384-322 A.C.). As classes sugeridas por Aristóteles são: substância (objetos físicos), quantidade (números), qualidade, relação, espaço, tempo, posição, estado, ação e afeição. A essa lista podem ser adicionados (conforme [ALL 95]) eventos, idéias, conceitos e planos. Duas classes muito importantes são ações e eventos. O estudo de eventos, coisas que acontecem no mundo, está presente em muitas teorias semânticas por terem relação com a maneira como são organizadas as interpretações das sentenças.

A ambigüidade lexical se dá quando uma única palavra possui mais de um sentido (ou, visto de outra maneira, apresenta mais de uma entrada em uma representação ontológica). A palavra *banco*, por exemplo, pode ora referir-se a instituição financeira, ora ao artefato utilizado para sentar-se.

Além da ambigüidade lexical, podemos ter a ambigüidade semântica estrutural, advinda de uma ambigüidade sintática. A frase *Cachorros e gatos felizes vivem na fazenda* é ambígua em relação ao alcance do adjetivo *felizes* (pode referir-se aos gatos apenas, ou aos cachorros e gatos). Outras formas de ambigüidade estrutural são puramente semânticas, e derivam de uma única estrutura sintática. Um caso comum é o escopo dos quantificadores. Por exemplo, a frase *Todos os garotos gostam de um*

cachorro pode significar que há um único cachorro de que todos os garotos gostam ou que cada garoto gosta de um cachorro diferente. Os dois significados apresentariam diferentes traduções em formas lógicas, que constituem o formalismo comumente utilizado para expressar a semântica das frases da linguagem natural.

A semântica lógica trata o significado através de uma especificação do domínio de conhecimento, de acordo com a teoria dos conjuntos. Para expressar o significado de expressões da linguagem natural em lógica, é preciso traduzir as expressões para uma linguagem lógica. Porém, para dar conta do alto poder de expressão da linguagem natural, é preciso recorrer a lógicas não clássicas. Essas lógicas incorporam noções mais complexas, como o tempo, por exemplo (que nas linguagens naturais manifestam-se nas conjugações verbais). Um outro exemplo é a noção de intensão: em lógica clássica, assume-se que o significado de um termo seja um referente ou um elemento do domínio, mas em linguagem natural, muitas vezes, utilizamos termos que não possuem um referente (ou extensão), mas que têm o seu significado associado a uma idéia ou conceito (a intensão). Um exemplo pode ser dado pela expressão *o primeiro homem a pisar em Marte*, outro exemplo clássico é a expressão *o unicórnio*. Muitos dos trabalhos em semântica de linguagem natural procuram estender a lógica para poder expressar noções mais complexas (como a lógica temporal e intensional, para os casos exemplificados acima).

Uma outra questão que recebe bastante atenção, no estudo do significado da linguagem natural, diz respeito a elementos utilizados para se fazer referência a objetos ou entidades do discurso ou domínio. Esses elementos são chamados expressões referenciais. Determinadas expressões têm como significado objetos ou entidades específicas no mundo. A semântica da lógica clássica, discutida anteriormente, é uma semântica referencial, ou extensional. Em linguagem natural também utilizamos alguns termos para indicar objetos do contexto ou evocar alguma entidade, e existem diversos meios de se fazer isso, cada um com diferente propósito. Podemos referenciar um objeto indeterminado (*um cachorro*), ou nos referirmos a um objeto com interpretação específica dentro de um contexto (*o cachorro do vizinho*). Outras expressões, sem um conteúdo semântico muito específico, podem ser utilizadas como apontadores para determinados elementos. Esses apontadores geralmente são utilizados para fazer referência a um elemento em evidência para os falantes, podendo ser um elemento introduzido anteriormente na fala ou discurso, ou ser um elemento presente no contexto físico da enunciação. Exemplos são os pronomes pessoais retos ou demonstrativos (*ele, ela, isto, aquilo*). Quando um pronome se refere a um elemento do discurso, esse elemento geralmente antecede o pronome e, nesse caso, diz-se que existe uma relação anafórica entre o pronome e o seu antecedente. Algumas vezes, porém, o objeto ou entidade sendo referenciado é especificado posteriormente no discurso. Nesse caso diz-se que existe uma relação catafórica entre o pronome e a expressão manifestada posteriormente.

A área de semântica é uma área de estudo mais nebulosa do que a sintaxe, por apresentar questões que são difíceis de tratar de maneira exata e completa. A questão do significado está ligada ao conhecimento de mundo e, além disso, ligada a questões mais obscuras como estados mentais e consciência. Para simplificar o estudo da semântica, costuma-se fazer determinados recortes teóricos que, conseqüentemente, limitam o poder de alcance das teorias propostas. Os estudos do significado que procuram integrar

outros fatores, como contexto e falantes, constituem uma outra área de estudo denominada pragmática.

Trabalhos bem conhecidos na área de pragmática dizem respeito ao acordo mútuo estabelecido entre os falantes na conversação [GRI 68, GRI 75], ou apresentam uma nova maneira de compreensão do significado da linguagem natural, que vê a linguagem como ação: a teoria dos atos de fala [AUS 62, SEA 69].

Os falantes da língua têm conhecimento sobre a forma de se comunicar e, muitas vezes, alguns dos princípios seguidos pelos falantes são independentes de linguagem. A seguir, são apresentados alguns exemplos [GRE 96]:

- Um assunto neutro e amigável para um encontro casual é o clima.
- Se você é o falante, você irá se referir a você usando a palavra "Eu" e não a palavra "você".
- Ao contar uma história a alguém você vai levar em consideração o que é familiar e o que não é familiar ao seu ouvinte.
- Se você está fornecendo alguma informação a alguém você irá fornecer informação suficiente e não informação adicional, além do solicitado.
- Se alguém faz uma pergunta, você dá uma resposta relevante ao tópico em questão.

A pergunta *Sobrou um pouco de café?*, por exemplo, pode ser interpretada pelo destinatário como uma solicitação do emissor para receber uma xícara de café tendo, assim, um significado de sentença diferenciado do significado de enunciação. Situações como estas ilustram a diferença entre o significado literal da linguagem e o significado da linguagem em uso, que é o objeto de estudo da pragmática.

É interessante observar que a pragmática não é apenas estudada por lingüistas, mas também por antropólogos, filósofos, psicólogos, sócio-lingüistas, psico-lingüistas e cientistas da computação. Para os filósofos, uma das preocupações é a habilidade dos falantes de fazer referência mútua, enquanto que, para os sócio-lingüistas, o interesse recai mais nas questões de interação comunicativa e no modo como estas podem ser influenciadas pela classe social, raça e gênero dos participantes.

A ciência da computação, mais especificamente a inteligência artificial distribuída, está interessada nos mecanismos interativos para modelagem de agentes e sociedades de agentes. Todo o estudo de comunicação entre agentes inteligentes tem como fundamento a teoria dos atos de fala de Austin e Searle [VER 97]. Diferentes tipos de enunciados têm diferentes efeitos nos estados perceptivos dos agentes e nos estados do mundo representados; de acordo com a teoria dos atos de fala, os enunciados realizam diferentes tipos de ação, conforme a classificação dada abaixo.

- Representativos: o falante comunica que acredita na verdade da expressão (por exemplo, através de asserção ou conclusão).
- Diretivos: o falante tem por intenção provocar o ouvinte a realizar uma ação (por exemplo, requisição, pergunta, ordem, proibição, permissão).
- Comissivos: o falante se compromete com a realização de uma ação no futuro (por exemplo, promessa, ameaça).

- Expressivos: o falante expressa um estado psicológico (por exemplo, agradecimento, pedido de desculpas).
- Declarações: têm como efeito imediato uma mudança de estado (por exemplo, uma declaração de guerra, a confirmação do batismo).

Classificações como estas são usadas de base para a construção de protocolos de comunicação entre os agentes.

Com essa discussão sobre semântica e pragmática, encerramos a apresentação dos níveis de estudo da linguagem. Diferentes aplicações em lingüística computacional irão privilegiar um ou outro aspecto, e diferentes soluções computacionais serão desenvolvidas de acordo. Algumas dessas soluções serão apresentadas na seqüência desse material.

3. O processamento da linguagem natural

A busca por entender os mecanismos da língua iniciou-se com os primeiros estudos de gramática na Grécia antiga, ganhou uma abordagem mais formal através dos estudos de Ferdinand de Saussure [*apud* FUC 92] e desenvolveu-se notoriamente através dos trabalhos de Frege [GEA 52], Noam Chomsky [CHO 57] e Richard Montague [DOW 81].

O interesse em dotar um sistema computacional com a capacidade de entender os objetivos do usuário em sua própria linguagem surgiu juntamente com os primeiros sistemas. Allan Turing, um dos maiores teóricos da computação, definia a inteligência dos computadores através da capacidade destes últimos em lidarem com a linguagem natural. A capacidade de processar linguagem natural, portanto, vem sendo pensada praticamente desde o advento dos computadores. Embora a máquina de Von Neumann tenha sido imaginada para aplicações numéricas, Turing já entendia o computador como um recurso com capacidades inteligentes, que o apoiaria em atividades como jogar xadrez ou teria, inclusive, habilidade para compreender e produzir linguagem natural.

Para Anton Nijholt em [NIJ 88], um propulsor considerável para a área da lingüística computacional foi a guerra fria. As aplicações de uso militar logo incluíram algoritmos de criptologia e os primeiros ensaios em tradução automática. Os projetos envolvendo tradução se multiplicaram até chegar-se, em 1966, a uma situação que contabilizava mais de 20 milhões de dólares gastos, com poucos resultados obtidos. Avaliada por um comitê nomeado para estudar o assunto, esta situação de custos exagerados mereceu um corte de financiamento por parte do governo americano. Com a redução de financiamento, passou a ser mais incentivada a pesquisa básica (como, por exemplo, a representação do conhecimento), dando-se menos crédito à pesquisa aplicada (como a tradução automática, as interfaces em linguagem natural etc). As pesquisas retomaram o rumo das aplicações nos anos 80, não sendo deixado de lado, entretanto, o trabalho com a teoria.

Com o retorno à ênfase nas aplicações, percebe-se também uma preocupação com a avaliação dos sistemas desenvolvidos e com a construção de sistemas com capacidade de processar a linguagem em larga escala (os primeiros sistemas, muitas vezes, demonstravam a aplicação de teorias em exemplos construídos, determinados e escolhidos com o propósito último de ilustrar o funcionamento do sistema). Para ilustrar os avanços da área, temos o auxílio dado à edição de documentos através da verificação

ortográfica e gramatical. Esses sistemas já atingiram um nível capaz de prover mais satisfação do que frustração ao usuário, apesar de ser ainda necessário continuar-se trabalhando para que melhorem. Para esta aplicação, temos um exemplo de uma ferramenta para o tratamento da língua portuguesa [NUN 00]. Outro exemplo de aplicação pode ser dado por sistemas de ditados (podemos ditar textos ao c omputador para que ele os escreva), desenvolvimento de interfaces baseaadas em fala (somos também capazes de ditar comandos ao nosso computador), e sintetizadores de fala (temos sistemas que poder "ler" textos escritos).

Os avanços na área, apesar de visíveis, enfrentam até hoje questões de difícil solução. Exemplos para ilustrar essa dificuldade podem ser obtidos observando-se a qualidade da saída fornecida pelos sistemas de tradução automática (uma análise detalhada é apresentada em [OLI 00]). Outro exemplo marcante é a dificuldade em conseguirmos respostas adequadas a perguntas, quando formuladas em linguagem natural, mesmo tendo computadores poderosos com acesso a grandes bases de dados (sejam elas textuais ou não).

Para lidar com os vários problemas, temos hoje, em nível mundial, uma comunidade científica e acadêmica em crescimento. Há muita pesquisa e trabalhos realizados principalmente para o Inglês, Espanhol, Alemão, Francês e Japonês. Encontramos, porém, carência de pesquisas, ferramentas, recursos lingüísticos e humanos para tratar computacionalmente a língua portuguesa. Todavia existem esforços para suprir essa carência. Um exemplo, em relação à formação de recursos humanos em nível de graduação, é a iniciativa das Faculdades da Universidade de Lisboa, Faculdade de Ciências (Departamentos de Informática e Matemática) e Faculdade de Letras (Departamento de Lingüística), que lançaram em 1994/1995 o curso de Licenciatura em Engenharia da Linguagem e do Conhecimento[2]. Exemplos de áreas de atividade econômica e aplicações que justificam a iniciativa são:

- Sistemas automáticos de indexação e categorização que classificam os documentos são fundamentais para lidar com a grande quantidade de informação produzida e manipulada em muitos setores de atividade. Seria também interessante que esses sistemas apresentassem a produção automática de resumos normalizados.
- É importante oferecer a usuários a possibilidade de acesso, em sua língua materna, a grandes bases de conhecimento sobre múltiplos domínios - transportes, seguros, meteorologia etc. Isto requer metodologias de organização da informação e sistemas de busca inteligente com interfaces em linguagem natural.
- Conversores de fala para texto e sistemas de apoio à tradução e ao diálogo multilíngüe ajudam a melhorar e a desenvolver a cooperação internacional.

Nesta seção iremos apresentar uma introdução aos princípios básicos que regem o desenvolvimento de sistemas de processamento da linguagem natural, procurando dar uma idéia do que está por trás de cada uma das aplicações que podemos presenciar hoje em dia. Um sistema para processar linguagem natural reúne, geralmente, alguns módulos organizados de acordo com a divisão vista nos estudos da lingüística. Cada

[2] http://www.lelc.f2s.com/mainframe.htm

uma das etapas do processamento exige um conhecimento de natureza diferenciada sobre a língua, e as soluções propostas irão variar de acordo com a natureza dos conhecimentos envolvidos. Para proporcionar ao leitor uma compreensão geral dos trabalhos realizados na área, iremos apresentar, nas próximas subseções, diferentes focos do processamento da língua natural, associados às etapas lingüísticas de processamento da língua.

3.1. Reconhecimento e síntese da fala

Avanços nos estudos sobre o reconhecimento da fala tornaram possível o desenvolvimento de sistemas que reconhecem as diversas palavras de uma língua. O reconhecimento pode ser de palavras isoladas, pertencentes a um vocabulário restrito, o que é útil para interfaces de alguns dispositivos. O reconhecimento também pode servir para ditar o nome de alguém para que o telefone efetue a chamada de um certo número, por exemplo; ou, pode ser útil reconhecer-se qualquer palavra de uma língua, através da fala contínua, o que é necessário aos sistemas de ditados, onde o usuário dita e o computador transcreve a fala em texto. Sistemas de síntese realizam o processo inverso: a partir de um texto escrito, o sistema faz a "leitura" em voz alta para o usuário.

Apesar de o estado da arte permitir a existência de produtos comerciais que realizem estas funções (*IBM Via Voice*, *Philips FreeSpeech*, são exemplos de sistemas para reconhecimento da fala, o *CMU Pronouncing Dictionary* é um exemplo de software para síntese) esta ainda é uma área que necessita de estudos e projetos como, por exemplo, a integração da tecnologia do reconhecimento de fala às interfaces de produtos de software ao usuário em geral. Outro problema que a penetração desse tipo de produto enfrenta, muitas vezes, é a necessidade de treinamento que o produto requer por parte do usuário. Longe de ter a facilidade de um *plug and play*, esses sistemas, uma vez adquiridos, devem-se adaptar à voz dos seus usuários. Para que um sistema seja independente de treinamento, ele deverá ser capaz de reconhecer as mesmas palavras sendo pronunciadas por diferentes vozes, com diferentes sotaques. Dado o estado atual da tecnologia para o reconhecimento da fala, para se oferecer um sistema independente de treinamento, o desenvolvedor desse sistema deveria realizar o treinamento do sistema em larga escala, suprindo uma ampla variação de pronúncias, o que, por sua vez, acrescentaria muito custo ao produto, tornando-o comercialmente inviável.

Uma área de estudo importante, hoje, é a que faz uso dos modelos probabilísticos de pronunciação e ortografia, e dos modelos probabilísticos de seqüências de sons produzidas pelos falantes da língua. Apenas para a palavra *the* da língua inglesa, por exemplo, já seria possível observar variações: *thee* e às vezes *thuh*. Note-se que estas variações não são exatamente originárias de regionalismos, mas sim da própria seqüência de palavras que sucedem ao *the* no discurso. Outras vezes, observamos palavras como *because* pronunciadas apenas como *cause*, por exemplo. Essas ambigüidades ou peculiaridades de pronúncia podem ser expressas através de regras que descrevem tais variações. Uma arquitetura comumente usada para levar em consideração tais variações [JUR 00] é a que utiliza o teorema de Bayes (ou método bayesiano) e o modelo de canal de comunicação com ruído. Essa arquitetura leva em conta probabilidades, produzindo um modelo que posteriormente poderá ser utilizado em associação com algoritmos de programação dinâmica, ou com o algoritmo de Viterbi, ou com o algoritmo da distância mínima de edição, entre outros. Também

existe a alternativa de associar-se o modelo probabilístico a um autômato de estados finitos, levando a um modelo de autômato com pesos associados.

Em relação à síntese de voz, os sistemas atuais ainda encontram dificuldades com relação aos aspectos prosódicos, ou seja, em reproduzir pronúncia, entonação e sotaque naturais. A preocupação com a prosódia, nesses sistemas, diz respeito aos aspectos da pronúncia de uma sentença que não estão descritos na seqüência de fones[3] derivados do léxico, mas se referem à produção de unidades lingüísticas maiores. Tais fenômenos geralmente são denominados fenômenos supras-segmentais, e envolvem a tonicidade, ritmo e pausas, pronúncia de combinações específicas de palavras, unidades de entonação, limites de frases e de sentenças e aspectos melódicos de sentenças.

3.2. Análise léxico-morfológica

3.2.1. Léxico

O léxico ou dicionário é uma estrutura fundamental para a maioria dos sistemas e aplicações. É a estrutura de dados contendo os itens lexicais e as informações correspondentes a estes itens. Em realidade, os itens que constituem as entradas em um léxico podem ser palavras isoladas (como *lua*, *mel*, *casa*, *modo*) ou composições de palavras as quais, reunidas, apresentam um significado específico (por exemplo, *lua de mel* ou *Casa de Cultura* ou *a grosso modo*). Entre as informações associadas aos itens lexicais, no léxico, encontra-se a categoria gramatical (*part-of-speech* ou *POS*) do item, além de valores para variáveis morfo-sintático-semânticas como gênero, número, grau, pessoa, tempo, modo, regência verbal ou nominal etc. Também são associadas ao item lexical, no léxico, uma ou mais representações ou descrições semânticas. Bem mais raramente, encontram-se associações a representações contextuais.

Na seção 2.3, comentou-se sobre a ambigüidade das palavras. O léxico irá representar, através das múltiplas descrições que podem estar associadas a uma entrada, os múltiplos sentidos de cada palavra ou item lexical.

Entre as estruturas mais utilizadas para reunir os itens de um léxico, duas se destacam: a estrutura de formas e a estrutura de bases. Um léxico pode conter todos os itens lexicais (sejam palavras ou unidades maiores que palavras) por extenso – neste caso será um dicionário de formas. A seguir apresentamos exemplos de entradas em um dicionário desse tipo:

a artigo feminino singular

determina um nome

a preposição

para, em direção a

a substantivo masculino singular normal

primeira letra do alfabeto

a pronome feminino singular 3ª pessoa

indica um referente feminino

[3] Unidades de composição fonológica da língua.

casa substantivo feminino singular normal

moradia, habitação, sede

casa verbo singular 3ª pessoa presente indicativo 1ª conjugação

contrair matrimônio

Outro modo de estruturar-se um léxico pressupõe colocar-se em evidência os morfemas que são os constituintes básicos das palavras (daí usando-se a denominação 'dicionário de bases'). Nesse caso, o léxico é constituído de unidades menores as quais concentram a capacidade de identificação de um determinado item (exemplo: *cas* para *casa*), e as diferentes formas desse item serão obtidas a partir de ligações com outras cadeias ou morfemas, através de regras de associação. Nota-se que as bases nem sempre correspondem exatamente aos morfemas básicos das palavras, nas aplicações. Freqüentemente se opta por inserir, no léxico, cadeias maiores, por conveniência do algoritmo que fará uso do léxico. Por exemplo, é comum considerar-se como bases cadeias que incluem prefixos, sem dissociá-los do morfema predominante, pois a gestão das composições do tipo prefixo+base pode onerar o sistema em termos de reconhecimento e geração de palavras (por exemplo, incluiríamos a base *preven* no léxico sem desvincular-se da mesma o prefixo *pré*, no caso do verbo *prevenir* ou do substantivo *prevenção*). As associações possíveis, para gerar novas formas a partir de uma certa base, podem ser representadas como uma rede de transição de estados. Para configurar as novas palavras a partir das bases, visando reduzir-se a multiplicidade de estados, podem ser criados modelos como, por exemplo, os associados a conjugações verbais para os verbos regulares, o plural em *s*, o diminutivo em *inh* etc.

Ao analisarmos estas duas alternativas de implementação do léxico, podemos efetuar uma breve comparação, que é compartilhada por Beardon em [BEA 91]. No caso do dicionário de formas, a disponibilidade de todos os itens lexicais diretamente no léxico facilita a busca às informações, tornando desnecessário um algoritmo que reconheça os itens a partir de seus constituintes. Já o modelo de léxico de bases e terminações é mais compacto e estruturado. Porém este modelo, embora elegante, exige, durante a etapa de análise, um algoritmo bem mais complexo, o qual deverá reconhecer individualmente os componentes de um item para, só então, produzir a análise.

A representação de um grande vocabulário através de um autômato de estados finitos é a alternativa proposta por Kowaltowski e Lucchesi e descrita em [KOV 93]. Essa foi a alternativa utilizada na implementação do amplo dicionário do Português Brasileiro e do corretor ortográfico hoje licenciados para o editor de textos *Microsoft Word*, e prevê que os vocábulos e as informações associadas sejam representados através de suas letras em um autômato finito. Os dicionários de bases e terminações podem ser entendidos como um caso específico de implementação de um transdutor de estados, com regras associadas a cada etapa de transformação.

3.2.2. Analisador léxico-morfológico

O analisador léxico-morfológico tradicionalmente decompõe a sentença em itens lexicais e realiza uma varredura, tratando item a item, e decompondo-os em seus morfemas. Verifica, a partir das informações armazenadas no léxico e nos morfemas, a estrutura, características e informações associadas a um determinado item, tais como gênero e número para substantivos, ou pessoa, número, modo e tempo, para os verbos,

por exemplo. Esta abordagem (decomposição da sentença em itens lexicais e aquisição de informações associadas a cada item) passa, atualmente, por um processo de avaliação: algumas aplicações não chegam a fazer uso de todos os resultados que podem ser obtidos ao recortar-se a sentença em itens, os itens em morfemas etc. Uma alternativa à análise léxico-morfológica tradicional vem sendo a etiquetagem automática de textos. Nesta seção apresentaremos cada uma destas alternativas.

A tarefa de análise, apesar de aparentemente simples, pode incluir problemas complexos. A morfologia nem sempre é sistemática, o que faz com que a decomposição em morfemas nem sempre seja clara. Outro fator de dificuldade são as variações ortográficas decorrentes da absorção dos morfemas (por exemplo, passa-se de *viagem* a *viajar*), o que leva a situações de substituição, acréscimo ou mesmo supressão de caracteres.

Para alguns autores como Bouillon [BOU 98], o analisador léxico-morfológico deve ainda ter capacidade gerativa. Isto significa que deve ocupar-se das possíveis combinações dos morfemas em palavras bem formadas: a geração de palavras.

3.2.3. Abordagem tradicional de análise

Dada uma determinada sentença, o analisador léxico-morfológico identifica os itens lexicais que a compõem e obtém, para cada um deles, as diferentes descrições correspondentes às entradas no léxico (isto é, categorias em que estes itens podem estar atuando e demais informações). A ambigüidade léxico-morfológica ocorre quando uma mesma palavra apresenta diversas categorias gramaticais. A palavra *a*, por exemplo, pode ser um artigo definido, uma preposição, um pronome, um substantivo (a letra *a*) etc. Em nível léxico-morfológico é importante que todas as formas possíveis de categorização sejam buscadas e informadas, independente de ocorrer ambigüidade. A ambigüidade será tratada em níveis mais avançados de análise.

Dizendo-se de outro modo, um programa que trata automaticamente a morfologia deve realizar a segmentação do texto de entrada e da sentença. Deve identificar o item lexical ou palavra desdobrando-o em morfemas, e associar corretamente as informações léxico-morfológicas a cada morfema, construindo assim o conjunto de informações léxico-morfológicas do item.

A implementação de analisadores léxico-morfológicos pode ser feita através de sistemas de índices, através de percurso em árvore, através de autômatos finitos, ou através de outras técnicas tais como a etiquetagem automática, bastante utilizada atualmente.

3.2.4. Etiquetagem (*POS tagging*)

O etiquetador gramatical (ou *pos tagger*) é um sistema responsável por identificar, em uma sentença, para cada um dos itens lexicais, a categoria a que este item pertence. Por exemplo, para a palavra *a* o analisador deverá decidir qual a categoria correta, de acordo com a posição que a palavra ocupa na frase. Neste caso, ao contrário do que se coloca quanto a oferecer 'todas as opções possíveis', deixando para uma próxima etapa a resolução das ambigüidades, o etiquetador está preparado justamente para tratar o texto de modo que este já sirva como entrada para aplicações, sem necessariamente passar por próximas etapas de processamento.

As etiquetas, ou partes do discurso, costumam incluir: substantivo (nome), verbo, pronome, preposição, advérbio, conjunção, particípio e artigo. Dependendo da aplicação para a qual servirá o texto etiquetado, o número de etiquetas pode variar: são 45 as etiquetas usadas no *Penn Treebank*[4] e 87 as usadas no *Brown corpus*. Essas são duas importantes coleções de textos em língua inglesa, etiquetados, disponíveis atualmente.

Embora sendo um processo que gera um resultado menos completo do que a análise léxico-morfológica convencional, ainda assim muitas informações são disponibilizadas, sobre a palavra (ou item lexical) bem como sobre seus vizinhos, e este vem se tornando um processo de análise muito difundido. O fato de saber que uma certa palavra é, por exemplo, um pronome possessivo, ajuda-nos a efetuar previsões sobre as palavras que a podem suceder, por exemplo, numa aplicação de reconhecimento da fala. O fato de saber que estamos lidando com um substantivo pode colocar em evidência o potencial dessa palavra para ser um dos indexadores do texto, em um ambiente de recuperação de informações.

A etiquetagem [JUR 00] é o processo de assinalamento de um marcador de classe gramatical (ou outro marcador ou 'etiqueta' de interesse) a cada palavra, num corpus. Esse processo corresponderia à 'tokenização', no processamento das linguagens de programação. A etiquetagem, como trata de linguagem natural, lida com um número bem maior de situações de ambigüidade. A entrada para a etiquetagem é uma cadeia de itens lexicais, e um conjunto específico de etiquetas; a saída é o conjunto de itens lexicais com a melhor etiqueta associada a cada item. O desafio do processo de etiquetagem reside exatamente em resolver as ambigüidades.

Os algoritmos para etiquetagem fundamentam-se em dois modelos mais conhecidos: os baseados em regras e os estocásticos. Os algoritmos baseados em regras, como o nome o diz, fazem uso de bases de regras para identificar a categoria de um certo item lexical. Neste caso, novas regras vão sendo integradas à base à medida que novas situações de uso do item vão sendo encontradas. Os algoritmos baseados em métodos estocásticos costumam resolver as ambigüidades através de um corpus de treino, marcado corretamente (muitas vezes através de esforço manual), calculando a probabilidade que uma certa palavra ou item lexical terá de receber uma certa etiqueta em um certo contexto. O etiquetador de Eric Brill [BRI 95], bastante conhecido na literatura, faz uso de uma combinação desses dois modelos.

---

[4] Um *Treebank* é um corpus de sentenças já corretamente analisadas e marcadas.

3.3. Análise sintática

Enquanto o analisador léxico-morfológico trabalha em nível de sentença, o analisador sintático trabalha em nível de frase (ou sintagma), e irá reconhecer uma seqüência de palavras como constituindo uma frase da língua ou não. Poderá também construir uma árvore de derivação, que explicita as relações entre as palavras que compõem a sentença. O analisador sintático faz uso do léxico, que reúne o conjunto de itens lexicais da língua, e de uma gramática, que define as regras de combinação dos itens na formação das frases.

3.3.1. Gramáticas e formalismos

A gramática utilizada para representar uma linguagem natural deve apresentar um bom balanço entre sua expressividade e o processo de reconhecimento. Chomsky [CHO 56] classificou as gramáticas em quatro tipos: tipo 3, regulares, tipo 2, livres de contexto, tipo 1, sensíveis ao contexto e tipo 0, sistemas de reescrita geral. As gramáticas do tipo 3, ou regulares, são as mais restritas, e por isso são as mais fáceis de serem reconhecidas. São, no entanto, insuficientes para expressar as regras de formação da linguagem natural. Gramáticas do tipo 2, livres de contexto, mais poderosas, permitem a representação de linguagens com um grau maior de complexidade; estas ainda apresentam problemas para expressar dependências, como é o caso da concordância verbal. O próximo nível de gramáticas, sensíveis ao contexto, resolve o problema das dependências, mas apresenta problemas de complexidade no reconhecimento. Decidir se uma sentença pertence a uma gramática sensível ao contexto é uma função exponencial sobre o tamanho da sentença, o que torna a implementação do procedimento de verificação uma questão complexa, do ponto de vista computacional.

A gramática adotada pode ser escrita através de diversos formalismos. Entre eles, podemos destacar [WOO 70] [FUC 93] [JUR 00] as redes de transição, [GAZ 82] gramáticas de constituintes imediatos (PSG ou *phrase structure grammar*), [GAZ 85] gramáticas de constituintes imediatos generalizadas (GPSG), [KAY 79] gramáticas de unificação funcional, [SHI 71] PATR-II e [POL 94] HPSG (*head-driven phrase-strutcture grammar*).

A decisão em relação ao melhor formalismo para representação da linguagem natural não tem ainda solução: as pesquisas têm proposto trabalhar em modelos que se situem em um nível intermediário entre as gramáticas livres de contexto e sensíveis ao contexto.

As gramáticas de constituintes imediatos (PSG), livres de contexto, apresentam a estrutura sintática das frases em termos de seus constituintes. Por exemplo, uma frase (F) é formada pelos constituintes: sintagma nominal (SN) e sintagma verbal (SV). O sintagma nominal é um agrupamento de palavras que tem como núcleo, ou elemento principal, um substantivo (Subst). O substantivo representa uma classe gramatical. No exemplo abaixo, são listados os substantivos *menino* e *chapéu*. O determinante (Det) compõe, junto com o substantivo, o sintagma nominal. Geralmente, um sintagma nominal possui uma formação mais complexa, podendo ter com constituinte uma oração (*o chapéu azul que eu comprei ontem*). O exemplo bastante simples, apresentado a seguir, ilustra uma gramática gerativa capaz de reconhecer a frase: *O menino usa o chapéu.*

F → SN, SV.
SN → Det, Subst.
SV → Verbo, SN.

Det → o
Subst → menino, chapéu
Verbo → usa

Esse formalismo gramatical oferece poder gerativo e capacidade computacional, e tem sido usado com sucesso em ciência da computação, na especificação de linguagens de programação, sendo que existem vários algoritmos eficientes para reconhecer as linguagens especificadas através do formalismo. Apresenta problemas, no entanto, em questões de concordância de gênero e número, que o formalismo não permite verificar. Se fôssemos incluir no léxico, como substantivo, os plurais, *os meninos*, ou o feminino, *menina*, por exemplo, as cadeias a seguir seriam aceitas como corretas. *O meninos usa os chapéu. O menina usa os chapéu.*

Det → o, os.
Subst → menino, menina, meninos, chapéu, chapéus.

O formalismo PATR-II permite verificar a concordância de gênero e número entre os constituintes da frase. Nesse formalismo as regras gramaticais informam sobre alguns traços sintáticos. Apresentamos a seguir a mesma gramática e léxico do exemplo acima, utilizando o formalismo PATR-II.

F → SN, SV
  <SN numero> = <SV numero>
  <SN pessoa> = <SV pessoa>
SN → Det, Subst
  <Det numero> = <Subst numero>
  <Det genero > = <Subst genero>
SV → Verbo, SN

o
  <categoria> = determinante
  <genero> = masc
  <numero> = sing
menino
  <categoria> = substantivo
  <genero> = masc
  <numero> = sing
chapéu
  <categoria> = substantivo
  <genero> = masc
  <numero> = sing
usa
  <categoria> = verbo
  <tempo> = pres
  <numero> = sing
  <pessoa> = 3
  <argumento 1> = SN
  <argumento 2> = SN

Durante a análise da frase, os valores dos traços sintáticos das palavras (obtidos do léxico) são utilizados para fixar os valores das variáveis associadas às regras da gramática, tornando possível a verificação de aspectos tais como concordância de gênero e número. De acordo com o especificado nesse formalismo, o número do sintagma nominal (SN) deverá ser o mesmo número do sintagma verbal (SV), não aceitando construções do tipo *Os meninos usa o chapéu*. Esses constituintes também devem concordar em pessoa, não permitindo, por exemplo, *Eu usa o chapéu*. Informação a respeito da subcategorização dos verbos também é fornecida. A subcategorização é a definição de argumentos que acompanham o verbo, e nesse exemplo é dada por sujeito (*o menino*) e objeto direto (*o chapéu*).

3.3.2. Métodos de análise

Tendo apresentado noções gerais sobre a gramática e os formalismos de representação, veremos a seguir diferentes métodos de análise sintática: os analisadores *top-down*, *bottom-up*, *left-corner* e tabular.

A linguagem de programação Prolog possui um formalismo para representação de gramáticas livres de contexto denominado DCG (*Definite Clause Grammar*), associado a um analisador *top-down* descendente recursivo. A conversão de regras da gramática vista anteriormente, de constituintes imediatos, em cláusulas Prolog, é muito simples. Para o exemplo precedente, temos:

```
f --> sn, sv.
n --> det, subst.
sv --> verbo, sn.

det --> [o].
subst --> [menino]; [chapéu].
verbo --> [usa].
```

Através dessa especificação, o interpretador Prolog irá reconhecer *o menino usa chapéu* como uma sentença válida da linguagem especificada, respondendo *sim* para uma consulta. O analisador irá procurar por um f, para obter o f irá procurar por um sn e um sv, para encontrar um sn irá procurar um det e um subst; quando obtém o det [o] e o subst [menino] ele completou um sn; passa então a procurar o sv, e assim por diante.

Para que o analisador, além de responder *sim* ou *não* sobre a validade da frase, gere a sua estrutura sintática, é preciso associar argumentos aos constituintes representados. Nesse caso, a consulta sobre a validade da sentença *o menino usa o chapéu* poderá reproduzir a estrutura a seguir:

```
(f(sn(det(o),subst(menino)),sv(v(usa),sn(det(o),subst(chapéu))) .
```

Para isso, a DCG deve ser modificada para:

```
f (f(SN,SV)) --> sn(SN), sv(SV).
sn(sn(Det,Subst) --> det(Det), subst(Subst).
sv(sv(V,SN)) --> verbo(V), sn(SN).

det(det(o)) --> [o].
subst(subst(menino)) --> [menino].
subst(subst(chapéu)) [chapéu].
```

```
verbo(verbo(usa)) --> [usa].
```

A possibilidade de inclusão de argumentos faz da DCG uma gramática mais poderosa do que a gramática de constituintes, permitindo tratar também a concordância:

```
sn(Numero,Genero) --> det (Numero,Genero), subst(Numero,Genero).

det(singular,masculino) --> [o].
det(singular,feminino) --> [a].
det(plural,masculino) --> [os].
det(plural,feminino) --> [as].
subst(singular, masculino) --> [menino], [chapéu].
```

Com esses argumentos, o analisador só aceitará os sintagmas nominais nos quais determinante e substantivo concordem em número e gênero.

A estrutura dos sintagmas verbais é variável de acordo com o verbo. Alguns verbos não exigem complementos além do sujeito, são os verbos intransitivos. Outros verbos só fazem sentido com a presença de um ou mais complementos. O tipo de complemento associado a cada verbo é denominado subcategorização. Note que os complementos podem variar em número e em tipo, alguns complementos são acompanhados de preposição (objeto indireto) ou não (objeto direto). O verbo *dar* por exemplo, refere-se a uma ação onde alguém, o sujeito, dá algo (objeto direto) a alguém (objeto indireto). O verbo *dar* requer, portanto, dois complementos: um sintagma preposicional, e um sintagma nominal. Para isso regras diferentes para sintagmas verbais devem ser adicionadas à gramática, onde o tipo de subcategorização associado a cada verbo pode ser representado:

```
sv --> v(1).
sv --> v(2), sn.
sv --> v(3), sn, sp.

v(1) --> [dorme].
v(2)--> [usa].
v(3)--> [deu].
```

Um problema com o analisador Prolog é que, por ser um analisador *top-down* da esquerda para a direita, ele entra em *loop* ao encontrar uma regra da seguinte forma

```
sn --> sn, conj, sn.
```

Essa regra diz que um sintagma nominal (SN) pode ser composto por dois SNs unidos por uma conjunção (conj), por exemplo, *o menino e a menina*, onde *e* é uma conjunção (conj --> [e].). O analisador irá proceder da seguinte forma: para analisar sn, irá procurar por um sn, e assim por diante. São regras recursivas à esquerda. Note que um *loop* como este pode estar distribuído em mais de uma regra:

```
sn --> conj_sns.
conj_sns --> sn, conj, sn.
```

É sabido que qualquer gramática recursiva à esquerda pode ser transformada em outra gramática que gera a mesma cadeia de palavras, mas não é recursiva à esquerda. O exemplo acima poderia ser reescrito como:

```
sn --> snx, conj, sn.
snx --> det, subst.
```

Porém, apesar de gerar a cadeia correta, essa escolha de regras irá produzir uma estrutura que não está de acordo com a gramática da língua, não há evidência lingüística de que exista uma diferença entre sn e snx. Por esse motivo, a transformação da gramática não é desejável nesse contexto. O que se faz é mudar para um tipo diferente de analisador, para evitar o *looping* e, ao mesmo tempo, manter correta a estrutura gerada.

O analisador *bottom-up* lê as palavras e tenta combiná-las em constituintes. Ao encontrar a palavra [o], reconhece-a com um det, encontra a próxima palavra [menino], é um subst, det e subst juntos formam um sn, e assim por diante. Pelo fato de que as ações desse analisador são disparadas por palavras, não há problemas em entrar em *looping* para regras recursivas à esquerda. Por outro lado, não pode lidar com constituintes vazios, cuja ocorrência é comum em Português (por exemplo: a supressão de pronomes: $[\ ]_{SN\text{-}1\ ^{a}\ pessoa}$ *uso o chapéu*.).

O analisador *left-corner* combina as estratégias *bottom-up* e *top-down*. Ao encontrar uma palavra, ele verifica que tipo de constituinte inicia com tal palavra e então faz o restante da análise de forma *top-down* para aquele constituinte. Dessa forma não há problemas em lidar com regras recursivas à esquerda.

Considere agora o seguinte conjunto de regras, usadas no reconhecimento de estruturas como *o menino de chapéu*:

```
sn --> det, subst.
sn --> det, subst, sp.
sp --> prep, sn.

prep --> [de]; [em]; [com].
```

O *parser* tenta a primeira regra sn, que funciona até uma parte, mas sobram palavras que não foram analisadas. Ele deve então retornar, esquecendo o trabalho de análise realizado até então, e tenta a próxima regra sn. Desse modo o analisador teve que percorrer o mesmo det e subst duas vezes, lembrando que a situação poderia ter ocorrido com estruturas mais complexas.

O analisador tabular (*chart parser*) tem condições de lembrar as subestruturas já analisadas e, se um retrocesso for necessário, a repetição pode ser evitada. Por exemplo, na primeira tentativa, o analisador tabular irá reconhecer [o] [menino] com um sn; na segunda tentativa, ao procurar por um sn no início de [o] [menino] [de] [chapéu], uma busca será feita em seus registros antes de usar mais uma vez a regra. Ao encontrar sn(det(o),subst(menino)) não será preciso iniciar uma nova análise de um sn.

Além das fontes específicas citadas ao longo do texto, a apresentação dessa seção baseou-se em [COV 94], para as questões relacionadas ao Prolog e DCG e [BAR 96] também foi consultado.

3.4. Análise semântica

3.4.1. O significado proposicional e a forma lógica

A análise sintática, estudada na seção anterior, permite verificar a boa formação das sentenças e frases de uma língua do ponto de vista estrutural, ou seja, levando em conta as combinações possíveis entre tipos de palavras. Exemplos de aplicação, decorrentes dessa análise, são os corretores gramaticais. Algumas combinações, no entanto, podem ser aceitáveis de um ponto de vista estritamente sintático e apresentar anomalias que são relacionadas ao conhecimento semântico. A inversão da sentença usada como exemplo na seção anterior pode ser usada para ilustrar essa situação: *o chapéu usa o menino*. Apesar de poder ser reconhecida como uma frase da língua (a estrutura sintática é idêntica a de sua inversão), pode-se perceber que uma sentença como essa apresenta uma dificuldade de interpretação.

Certas aplicações necessitam lidar com a interpretação das frases bem formadas, não bastando o conhecimento da estrutura, mas sendo necessário o conhecimento do significado dessas construções. Podemos querer que respostas sejam dadas a sentenças ou orações expressas em língua natural as quais, por exemplo, provoquem um movimento no braço de um robô. Ou podemos querer extrair conhecimentos sobre o tema ‘indústria automotora’ a partir de uma base de dados textuais.

Num tratamento automático, a análise semântica [FUC 93] consiste em associar, a uma seqüência de marcadores lingüísticos, uma ‘representação interna’, entendida como a representação do significado desta sentença. A seqüência de marcadores aqui citada geralmente é a proveniente da análise sintática. Não obstante, para certas aplicações bem específicas, a representação do significado pode ser construída sem necessidade de uma análise sintática preliminar ou conjunta.

O nível semântico de conhecimento é bem mais difícil de descrever que os precedentes (léxico-morfológico, sintático). As aplicações bem sucedidas normalmente se restringem a um domínio circunscrito de conhecimento.

Uma das maneiras de abordar a semântica da linguagem natural é através da especificação do ‘significado proposicional’. Esse está diretamente ligado às formas lingüísticas presentes na sentença e difere do ‘significado pragmático’, ou sentido, que a sentença assume num certo contexto. Este último é objeto da análise pragmática. Sob o enfoque do significado proposicional, a análise semântica envolve a tradução de uma sentença em linguagem natural para uma expressão em linguagem formal, que é bem discriminada. Em comparação às linguagens naturais, as linguagens formais, tais como a linguagem lógica, apresentam uma semântica bem definida. Por isso, existe uma grande influência da lógica nos estudos da semântica computacional da linguagem natural. De acordo com a estrutura sintática de uma sentença, é possível estabelecer uma representação lógica correspondente, onde o verbo indica uma relação entre os argumentos expressos por sujeito e complemento verbal (objeto direto ou indireto). Aspectos da pragmática (como os aspectos contextuais, atos de fala etc) incidirão sobre essa representação.

Um exemplo de trabalho mais específico, desenvolvido em semântica lógica da linguagem natural, é a tradução de quantificadores da linguagem natural para os da linguagem lógica (quantificador existencial, quantificador universal). Para os

quantificadores, encontramos problemas de ambigüidade muitas vezes relacionados à definição de escopo. A frase *Todo homem ama uma mulher* pode ser interpretada através de duas formas lógicas distintas: em uma delas, existe uma única mulher amada por todos os homens; na segunda, cada homem ama uma mulher e estas podem ser diferentes.

A lógica foi, muitas vezes, desafiada pela linguagem natural. Isto é, para expressar a semântica da linguagem natural de modo mais fiel, propostas de alteração à lógica foram apresentadas. Um exemplo dessa situação envolve os quantificadores: enquanto que a lógica clássica tem dois quantificadores, o *para todo* ($\forall$) e o *existe* ($\exists$), em linguagem natural temos outros quantificadores, com significados diferenciados como, por exemplo, *muitos*, *poucos*, *nenhum*, *pelo menos x*, *no máximo x* etc. Para que possam ser representados, estes quantificadores exigem extensões da lógica clássica.

3.4.2. Fenômenos semânticos

As seqüências cujo significado o analisador semântico deve descrever, normalmente, se compõem de itens lexicais, analisados do ponto de vista léxico-morfológico e agrupados em estruturas por um processo de análise sintática. Essas organizações permitem desdobrar a semântica em estudos de duas naturezas distintas: uma semântica dita lexical, e uma semântica dita gramatical. A semântica lexical, ou semântica das palavras, está mais claramente associada às categorias de palavras como verbos, substantivos e adjetivos (também conhecidos como ‘palavras cheias’). Já as preposições e artigos (conhecidos como ‘palavras vazias’) estão mais associados à semântica gramatical. É costume associar-se, às palavras cheias, uma representação conceitual que descreva seu significado.

Alguns fenômenos ditos semânticos já são bastante estudados, como é o caso da ambigüidade proveniente da polissemia. Outras situações de interesse dizem respeito às relações interproposicionais (ou seja, entre frases distintas), às relações de referência, determinação e temporalidade. Esses fenômenos podem envolver conhecimentos adicionais além do conhecimento semântico, sendo estudados em um nível pragmático de tratamento.

3.4.3. Semântica lexical

A descrição semântica pode ser obtida por diferentes métodos de representação. Por exemplo, traços semânticos, como cor ou gênero, podem associar, aos itens lexicais, um certo número de características. Outro modo de fazê-lo seria através do uso de traços binários (por exemplo: para ‘uso’ teríamos ‘usado’ ou ‘novo’, o que pode ser representado por + uso ou - uso). Vamos explorar agora, em maior detalhe, questões relacionadas com a semântica lexical.

A representação de informação semântica pode estar presente no léxico (o que a torna útil, inclusive, à análise sintática). Um exemplo desse tipo de informação é dado pelas restrições de seleção. Na interpretação de linguagem natural, essas restrições auxiliam na eliminação de ambigüidade léxica. Voltando ao exemplo da palavra *banco*, instituição financeira e artefato usado para sentar, com base nas restrições, o sistema pode ser capaz de identificar o significado correto para *banco* em *O banco me forneceu um empréstimo.* Apresentamos abaixo as estruturas que seriam as entradas lexicais, com restrições de seleção associadas.

banco → [- objeto físico], [+ instituição]
banco → [+ objeto físico], [+ artefato]

O léxico pode também obedecer a regras de redundância e postulados semânticos como, por exemplo:

[+ humano] → [+ animado]
[+ humano] → [- abstrato]
possui(x,y) → pertence-a(y,x)

Os traços semânticos informados no léxico podem ser ainda utilizados para restringir as possibilidades de combinações entre as palavras, identificando incoerências semânticas. As classes semânticas utilizadas em restrições podem ser organizadas hierarquicamente em ontologias. Uma ontologia é um modelo 'extra-lingüístico' de conhecimento. Contém informações extra-lingüísticas organizadas em uma rede de conceitos, com definições de objetos, relações e propriedades, e as relações entre estes. As ontologias apresentam a modelagem do conhecimento associado a um certo domínio em particular. Por exemplo, na análise de *vários soldados atiraram nos homens e alguns caíram* um sistema que disponha unicamente de informações semânticas isoladas sobre as palavras não poderá identificar corretamente qual o antecedente de *alguns* (*soldados* ou *homens*?). Uma ontologia proveria o conhecimento sobre a relação de causalidade entre *atirar* e *cair*, o que permitiria identificar que os homens, e não os soldados, teriam caído (este exemplo é discutido em detalhe em [BOU 98]). Outra área que tem recebido bastante atenção é a do uso de ontologias para busca de informação. O trabalho sendo feito através do projeto SEMA, na PUCRS, tem como foco uma abordagem dessa natureza [GON 00]. As relações entre palavras no português vêm sendo estudadas de modo a representarmos ligações que sejam importantes, no momento de indexar a informação contida em documentos escritos, e no momento de recuperá-la. Neste caso, entretanto, já estamos lidando com conhecimentos que transcendem os itens lexicais isolados, o que aponta para uma semântica gramatical, e não apenas lexical.

3.4.4. Semântica gramatical

Uma análise semântica que se reduza à semântica lexical é insuficiente. A semântica gramatical procura descrever o significado da frase traduzindo-a em uma estrutura que interprete as relações sintáticas entre os itens lexicais. As relações podem ser representadas, por exemplo, através de uma estrutura associada a um certo verbo. Na sentença *João chutou a bola* observamos a mudança de estado de um objeto por força da ação de um sujeito. Várias outras sentenças poderão seguir a este 'padrão semântico'. Observe, por exemplo, *Maria bateu a porta* ou *Silvia fechou o livro*.

Uma forma de representar essas relações é a baseada em argumentos: cada proposição pode ser representada como uma relação predicativa constituída de um predicado, de seus argumentos e de eventuais modificadores. Essa representação é usada, por exemplo, em sistemas de tradução automática.

Outra forma de representar as relações semânticas é proposta através das gramáticas de casos. A base dessa abordagem é que um pequeno número de 'casos semânticos' (por exemplo: agente, objeto, instrumento etc) permitiria dar conta de todas

as construções, e seria possível estabelecer uma correspondência entre casos semânticos e funções sintáticas.

3.4.5. Formalismos de representação semântica

Na construção das representações semânticas, dois grupos de formalismos são mais usados: as estruturas do tipo ‘atributo-valor’ e os formalismos lógicos. Ambos os grupos já foram trabalhados ao longo deste texto. Os pares ‘atributo-valor’ permitem implementar, por exemplo, os traços semânticos mencionados ao longo da subseção sobre semântica lexical. O valor associado a um atributo pode ser simples (por exemplo, pode ser binário) ou complexo, ligando uma unidade semântica a outra, e produzindo uma estrutura de grafo. O primeiro dos exemplos a seguir mostra uma representação em pares ‘atributo-valor’ simples enquanto que o segundo exemplo (Figura 1) mostra uma estrutura de grafo a qual representaria informações equivalentes ao primeiro exemplo:

Canário

Tipo-de: pássaro

Cor: amarelo

Propriedade: assobiar

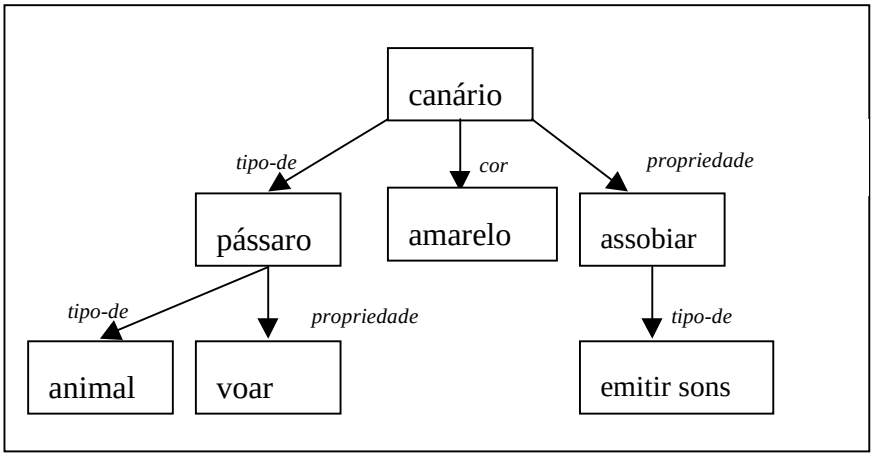

Figura 1: Representação semântica em forma de grafo

Os grafos também constituem a forma de representação utilizada no sistema de primitivas proposto por Schank [apud BEA 91], denominadas ‘primitivas de dependência conceitual’, que servem para representar conhecimentos semânticos. Para Schank, todas as ações podem ser decompostas em 11 conceitos de base, ou primitivas, tais como ‘aplicar uma força a um objeto’, ‘mudar a posição de um objeto’, ‘produzir um som’, ‘transferir informações de um indivíduo a outro’ etc. A estas primitivas Schank associa atributos, e então uma frase é representada por uma combinação de primitivas e atributos. Outras correntes teóricas também seguem uma representação em primitivas, entre as quais podemos citar a semântica preferencial de Wilks, e os grafos conceituais de Sowa. O interesse nesses sistemas está na capacidade de realizarem uma decomposição semântica.

Os formalismos lógicos são a outra grande vertente utilizada para a representação semântica. Podemos representar, através de fórmulas lógicas, os

conhecimentos lexicais. Nesse caso, a cada conceito ou significado, corresponderá um predicado com um número fixo de argumentos. Por exemplo, *dar (X, Y, Z)* pode representar a ação do agente X de dar o objeto Y a Z. Esse mecanismo poderá permitir uma série de inferências, porém pode ser necessário, por vezes, bloquear algumas heranças. Por exemplo, o avestruz é uma ave mas não voa. O poder de expressão da linguagem natural obriga à busca por novos formalismos lógicos para uma representação adequada, e assim temos, por exemplo, as lógicas temporais, que permitem expressar situações condicionais futuras, como *existirá um momento em que ela se dará conta da necessidade de poupar energia*.

Observamos que os formalismos lógicos são adequados à representação do sentido da frase ou sentença, já que permitem uma interpretação natural de uma asserção por uma forma predicativa.

3.4.6. Construção de representações semânticas

Para demonstrar a construção de uma representação semântica através de uma linguagem lógica podemos recorrer à DCG do Prolog vista na seção anterior. Para propósitos de ilustração, apenas, usaremos um subconjunto bastante simplificado do português onde os únicos sintagmas nominais são os nomes (rex, felix) e a seguinte sintaxe:

```
f--> sn, sv.
sv--> v(1).
sv-->v(2), sn.

v(1)-->[dorme].
v(2)-->[persegue].
sn--> [rex].
sn-->[felix].
```

Em lógica, uma representação semântica para nomes pode ser dada por uma constante individual, ou seja a constante rex para o indivíduo Rex, felix para Felix. As sentenças em linguagem natural serão representadas por sentenças da lógica de predicados de primeira ordem: Rex persegue Felix, persegue(rex,felix). Para representar os verbos isoladamente podemos utilizar expressões lambda (λ)(uma fórmula com falta de um argumento):

```
dorme = (λx) dorme(x)
```

onde λx indica que o valor de x deve ser fornecido. Quando dois argumentos são necessários, representa-se com uma expressão lambda dentro da outra:

```
persegue = (λy) (λx) persegue(x,y)
```

que significa, forneça-me o valor de y, por exemplo felix, e uma outra expressão lambda será retornada e que necessita um argumento para x tal que λx persegue(x,felix). A composição da representação semântica de uma frase será dada pela combinação da representação das palavras, por exemplo rex ao combinar-se com (λx)dorme(x) resultará em dorme(rex). Em Prolog, representamos o operador lambda com ^, e a DCG modificada para resultar em uma representação semântica é apresentada a seguir:

```
f(Predicado)--> sn(Sujeito), sv(Sujeito^Predicado).
sv(Sujeito^Predicado)--> v(Sujeito^Predicado).
```

```
sv(Sujeito^Predicado)-->v(Objeto^(Sujeito^Predicado)), sn(Objeto).

sn(rex) --> [rex].
sn(felix) --> [felix].
v(X^dorme(X)) --> [dorme].
v(Y^(X^persegue(X,Y))) --> [persegue].
```

O analisador irá responder a consultas da seguinte forma:

```
?-f(Semantica,[rex,persegue,felix],[]).

Semantica = persegue(rex,felix)
```

A Figura 2 abaixo mostra como essa construção é realizada. O significado é construído ao percorrer-se o caminho até o topo da árvore.

O exemplo simples apresentado acima permite ilustrar as idéias básicas envolvidas na tradução de linguagem natural para a linguagem lógica. Estudos mais avançados dessa questão incluem as representações semânticas para os quantificadores da linguagem natural. Esses devem ser traduzidos para os quantificadores lógicos. Essa tradução apresenta um grande número de problemas e por isso constitui uma área de estudos específica da semântica computacional, um trabalho clássico nessa área é [COO 83].

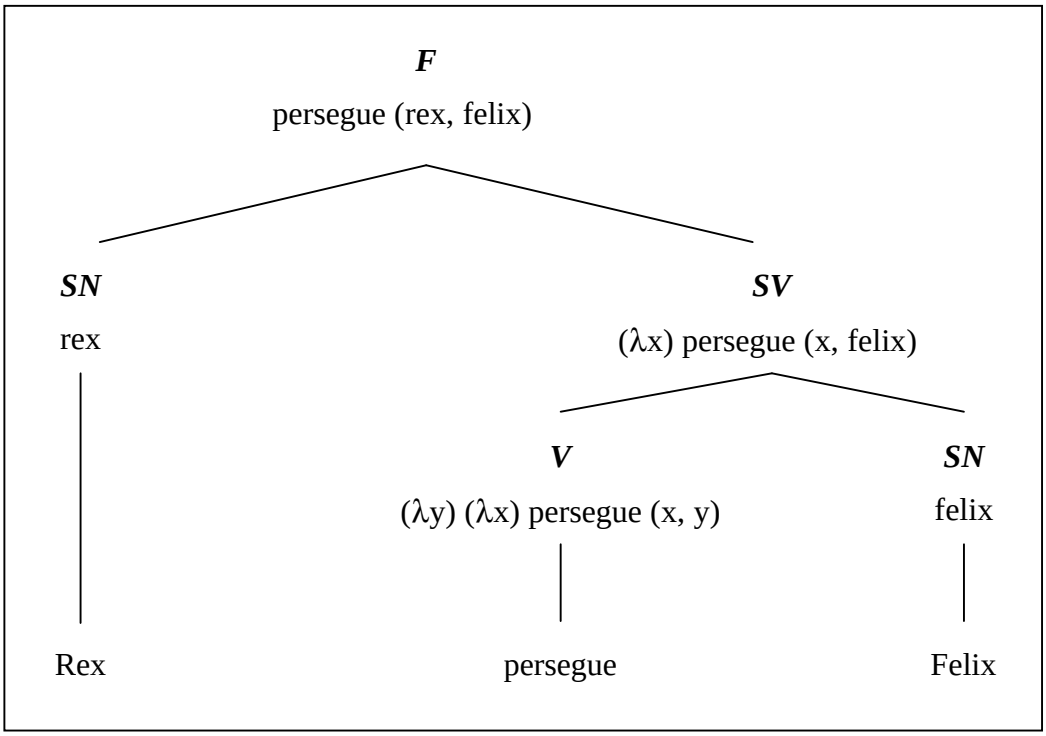

Figura 2: Construção da representação semântica

3.4.7. Princípios da análise semântica

Podemos considerar que a análise semântica seja realizada completamente em separado da análise sintática, e nesse caso a entrada do analisador semântico seria composta das árvores sintáticas associadas à sentença. Entretanto, a maior parte dos sistemas reúne as fases de análise sintática e semântica.

Para alguns teóricos, a análise semântica pode ser composicional, isto é, opera nodo a nodo a partir da árvore sintática. Para outros teóoricos, a análise semântica deve mesmo guiar a análise sintática, oferecendo, por exemplo, primitivas conceituais quando certas palavras são detectadas (seria o caso especial dos verbos). Então o analisador sintático-semântico procuraria 'preencher' diretamente os papéis semânticos previstos por uma certa primitiva, apoiado nos mecanismos da sintaxe.

Nota-se, entretanto, que muito ainda há por ser feito no domínio da análise semântica, tanto no que se refere à adequação dos formalismos, como no que se refere à representação do conhecimento semântico propriamente dito.

3.5. Análise pragmática

3.5.1. Pragmática e compreensão

A análise pragmática se refere à obtenção do significado 'não literal' de uma sentença. Ou seja, o significado completo, tal como o ser humano o percebe ao ler ou ouvir uma sentença, contém elementos que não estão representados unicamente nas unidades e nas relações semânticas. Além do conteúdo dito 'literal', há a necessidade de ligar as frases entre si, de modo a construir um todo coerente, e de interpretar a mensagem transmitida, de acordo com a situação e com as condições do enunciado. Por exemplo, examinemos a sentença: *o professor disse que duas semanas são o tempo necessário para resolver este problema*. Para uma compreensão literal, poderíamos recorrer aos mecanismos de representação expostos até aqui, e não teríamos dificuldades. Mesmo uma tradução poderia ser feita a partir dessa sentença, para um outro idioma. Entretanto, uma compreensão aprofundada exigiria saber a que problema se refere o professor, já que o problema deve ter sido a própria razão da formulação dessa sentença.

Dois pontos focais da pragmática são: as relações entre frases (para construir uma representação do texto, a representação de cada nova frase se apóia na precedente) e o contexto (a situação e condições em que ocorre o enunciado).
À medida que vão sendo enunciadas, as sentenças criam um universo de referência, que se une ao já existente. A própria vizinhança das sentenças ou dos itens lexicais também constitui um elemento importante na sua interpretação: o co-texto.
Assim, alguns novos fenômenos passam a ser estudados, como fenômenos pragmático-textuais. Inserem-se nessa categoria as relações anafóricas, co-referência, determinação, foco ou tema, dêiticos e elipse.

3.5.2. Questões pragmáticas

A pragmática relaciona a língua e seu uso. Esse uso inclui uma abrangência maior do que, simplesmente, sentenças isoladas, e a unidade de estudo passa a ser o discurso. Entende-se por discurso o texto ou a fala, compostos de várias unidades menores, que seriam as sentenças. Em nível de análise do discurso iremos encontrar algoritmos para resolução de referência, compreensão de diálogos e modelos de interpretação de textos em geral.

Alguns sistemas de processamento da linguagem natural possuem um mecanismo de inferências. Nesse caso, a ontologia pode colaborar para fornecer informações implícitas no texto. Por exemplo, para *Ana comprou um apartamento* o

sistema pode inferir que, antes da compra, Ana tinha o dinheiro correspondente ao preço do apartamento e que, agora, Ana possui um apartamento.

Outro problema é o da identificação de um significado, em determinado uso, para palavras polissêmicas (palavras que apresentam mais de um significado). Essa questão pode ser analisada sob a ótica dos contextos lingüístico e extra-lingüístico (vistos na seção 2.3).

O problema da resolução de anáforas diz respeito a encontrar os antecedentes que participam no processo de interpretação de determinadas expressões, por exemplo, os pronomes, sendo um tema na área limítrofe entre semântica e pragmática. A interpretação de um pronome (*ele*, *ela*, *isso*, *essa* etc) é relativa ao contexto de uso e, geralmente, em textos escritos, é relativa ao contexto lingüístico (isto é, é baseada em um antecedente lingüístico). Diversos algoritmos foram propostos para fazer a identificação do antecedente anafórico de pronomes, e novos modelos teóricos foram desenvolvidos para dar conta de questões relacionadas. Um exemplo é a teoria de representação do discurso (DRT *discourse representation theory*) [KAM 93]. Outro é o da teoria de *Centering* (*apud* [JUR 00]). O grupo de pesquisas em processamento da linguagem natural da PUCRS vem atuando na área da resolução de pronomes fazendo uso da teoria de *Centering* para resolução dos pronomes pessoais em português [PET 99] e também de abordagens baseadas em regras na resolução de possessivos e demonstrativos [SAT 01].

Alguns trabalhos levam em consideração alguns tipos particulares de expressões, por exemplo, as descrições definidas (aqueles sintagmas que iniciam por artigo definido). Um exemplo de estudo desenvolvido para tratar da resolução de co-referência do artigo definido pode ser o dado em [VIE 98]. Estudos similares estão sendo produzidos, mais recentemente, para a língua portuguesa [VIE 00].

Outros trabalhos, principalmente os sistemas participantes da série de conferências em compreensão de mensagens (*MUC Message Understanding Conference*) têm tratado da questão da co-referência de maneira mais geral. Nesse outro enfoque, o problema é o de reconhecer as diversas expressões cuja interpretação aponta ao mesmo referente.

Os significados implícitos também são um ponto de interesse na questão pragmática. É nessa dimensão que procura atuar a teoria dos atos de fala [Austin apud FUC 92]. Por exemplo, quando alguém diz *eu declaro a conferência aberta*, na verdade a carga de significado vai além da simples declaração: o ato (de abertura da conferência) é realizado diretamente ao serem proferidas estas palavras. Existem atos de fala diretos (como em *eu prometo ir*) ou indiretos, que exigem uma reconstrução por um mecanismo de inferência (como em *está fazendo frio aqui*, onde o falante pode estar solicitando que seja fechada a janela). A dimensão implícita exige o conhecimento das pressuposições. Por exemplo, em *tua irmã ainda toca piano?* Pressupõe-se que o interlocutor tenha uma irmã e que ela já tenha tocado piano.

Todas estas questões são ainda objeto de estudo de modo a prover mecanismos de representação e de inferência adequados, sendo raramente tratadas pelos sistemas de processamento.

3.5.3. Representação do discurso

Modelos de representação do discurso são necessários quando se tem por objetivo a construção de uma base de conhecimento de uma entrada textual em linguagem natural. Um primeiro problema que surge é a maneira como devemos considerar e representar as entidades mencionadas. Para a frase *João tem um cachorro,* não podemos simplesmente fazer uso direto da representação em Prolog a seguir:

```
cachorro(X).
possui(joão, X).
```

Essa representação corresponde ao fato "qualquer coisa é um cachorro" e "João possui qualquer coisa", e daria uma resposta afirmativa a uma consulta do tipo "?-possui(joão, new_york)". É preciso reconhecer o cachorro mencionado como um referente do discurso (algo sobre o qual podemos falar) e dar a ele um nome único (um identificador), por exemplo, ent_disc(123), e então a representação em Prolog para a frase acima é dada por:

```
cachorro(ent_disc(123)).
possui(joão,ent_disc(123)).
```

É comum em discursos em linguagem natural, o uso de anáforas (pronomes) para fazer referência a entidades previamente mencionadas. Para uma anáfora ser compreendida ela deve ser identificada com um referente do discurso previamente determinado. Esse processo é chamado de resolução de referência anafórica e envolve procurar por um antecedente, isto é, a referência anterior feita no discurso para essa mesma entidade. Os pronomes possuem alguma informação que pode ser utilizada no processo de identificação de um antecedente, tais como gênero e número. Mostraremos, para fins ilustrativos, um algoritmo bastante simples e genérico para resolução anafórica:

- Mantenha uma lista de referentes de discurso, ordenados por ordem de ocorrência no discurso, e marque cada um deles com condições de restrições: gênero (masculino, feminino), número(singular, plural), e ontológicas(animado, inanimado).
- Ao encontrar uma expressão anafórica, procure entre os referentes da lista, o mais recente que satisfaça as condições de restrição.

3.6. Considerações sobre abordagens híbridas: simbólicas e estatísticas

Além do processamento simbólico, tomado como base até aqui para explicar principalmente a análise sintática, é possível encontrar aplicações de processamento que fazem uso de outros métodos de análise. Particularmente, podemos observar o destaque que vem sendo proporcionado aos métodos estatísticos ou aos métodos híbridos.

As abordagens de pesquisa em lingüística computacional, durante um certo tempo, costumavam ser bem definidas em relação aos seus métodos. De um lado, tem-se a pesquisa de teorias motivadas pelos aspectos cognitivos da linguagem, de acordo com a tradição em lingüística gerativa. De outro lado, encontramos as abordagens motivadas por achados empíricos, baseados em coleções de dados lingüísticos ocorridos naturalmente. Os maiores influentes dessas duas correntes de abordagens computacionais à linguagem foram Chomsky [CHO 57] e Shannon-Weaver [SHA 49].

A maioria dos trabalhos em lingüística computacional desenvolveu-se de acordo com a perspectiva teórica da gramática gerativa (hostis aos métodos quantitativos), enquanto a comunidade voltada ao processamento da fala seguia os métodos estatísticos da teoria de informação (hostis a teorias lingüísticas). Durante algum tempo as duas áreas desenvolveram-se independentemente, sem diálogos. Nas décadas de 80 e 90, começaram a aparecer trabalhos na identificação de categorias sintáticas das palavras em uma frase, ou na resolução de ambigüidade de sintagmas preposicionais, com base nos mesmos métodos estatísticos já tradicionais em trabalhos de processamento de fala, e com sucesso. O conhecimento necessário para a solução de problemas começou a ser induzido pela análise de grandes corpora, ao invés de ser construído manualmente na forma de regras simbólicas. Desde então, cada uma das comunidades tem aceitado melhor a idéia de que, para se atingir os objetivos de cada área, pode ser necessário o conhecimento amadurecido pela outra.

Abordagens quantitativas passaram a adicionar robustez e abrangência a sistemas simbólicos de processamento de linguagem natural, os quais, até então, apresentavam alcance limitado, permitindo-lhes, por exemplo, a aquisição automática ou semi-automática de conhecimento lexical (terminologia, nomes próprios, eqüivalência em traduções). As abordagens quantitativas, por outro lado, careciam de informações sobre a natureza lingüística dos dados. Atualmente a convergência de abordagens é característica da área, e é reconhecida a necessidade de construção de sistemas efetivos e robustos que possam ser avaliados. Uma boa coleção sobre trabalhos apresentando soluções híbridas é apresentada em [KLA 96].

Pudemos observar na seção 3.2 o modo como, na análise léxico-morfológica, vêm sendo empregados, atualmente, métodos de etiquetagem automática, os *part-of-speech taggers*. Em especial, observamos que a etiquetagem é um processo de marcação que associa *taggers*, ou etiquetas, às palavras ou itens lexicais do texto de entrada. Esse processo de etiquetagem pode ser estendido para a marcação de informações mais completas sobre a estrutura sintática dos textos, que incluem a identificação, por exemplo, dos constituintes como sintagma nominal, sintagma verbal, sintagma preposicional etc. Ora, esta marcação irá suprir, em muitas aplicações, as funções de uma análise sintática.Nesse processo são utilizados os modelos de Markov [JUR 00], ou *Markov Models* (MM), os quais servem à modelagem de uma seqüência de eventos. Esses modelos trabalham com a ordem das palavras na sentença, podendo utilizar a ordem visível das palavras (*Visible Markov Models* ou VMM) ou a ordem “oculta” dessas palavras (*Hidden Markov Models* ou HMM), ou seja, um nível de abstração mais alto com relação à possível seqüência das palavras na sentença. No caso dos HMM, que são o modelo mais utilizado, esse nível adicional de abstração permite inserir estruturas adicionais, para visualizar a ordem das categorias das palavras.

O processo de marcação visa encontrar a seqüência mais provável de marcas, rótulos ou etiquetas que correspondam a uma dada seqüência de palavras. Para computar a seqüência de estados mais provável, normalmente é utilizado o algoritmo de Viterbi (descrito em detalhe em [GAS 00]).

O processo assim organizado prevê a existência de pelo menos dois corpora: um corpus de treino, marcado e revisado, a partir do qual o etiquetador irá ‘aprender’ regras, e o corpus de textos a serem analisados.

A eficiência de um sistema dessa natureza depende [MAN 99] de fatores como: quantidade de dados de treino disponíveis (quanto mais dados de treino, melhor); quantidade de etiquetas (maior a quantidade de etiquetas ou rótulos, mais específico o resultado, porém maior a possibilidade de ambigüidade); similaridades e diferenças entre corpus de treino e corpus de teste (se o corpus a ser etiquetado difere muito, em estilo ou gênero, do corpus utilizado para treinar o etiquetador, a precisão da marcação irá degradar); existência de palavras ou construções desconhecidas (a presença de palavras ou construções desconhecidas piora consideravelmente a qualidade dos resultados).

As equipes do GLINT, da Universidade Nova de Lisboa, coordenada pelo Prof. José Gabriel Pereira Lopes, em Portugal, e do NILC, sediada na Universidade Federal de São Carlos em São Paulo, coordenada pela Profª Maria das Graças Volpe Nunes, trabalham intensamente com a abordagem estatística e textos etiquetados. O grupo coordenado pelo Prof. Eckard Bick, na Universidade de Ahrus, na Dinamarca, vem trabalhando nos últimos 5 anos, através do projeto *Visual Interactive Syntax Learning*, com análise sintática de várias línguas, entre elas o português. Atualmente, pela Internet[5], é possível executar a análise sintática de textos da língua portuguesa.

4. Aplicações e desenvolvimento

Nesta seção serão discutidas diversas aplicações decorrentes do estudo e desenvolvimento da área de lingüística computacional.

4.1. Reconhecedores e sintetizadores da fala

Sistemas reconhecedores da fala têm sido utilizados para fins de ditado, onde o sistema faz a transcrição da fala em texto; em interfaces de comando por voz, por exemplo, para comandar o seu editor de texto ou navegar na Internet falando com o computador; ou em acesso a serviços automatizados de informação por telefone. Exemplos de produtos comerciais disponíveis no mercado são o *IBM Via Voice* e o *Philips FreeSpeech*, que apresentam versões para o reconhecimento da língua portuguesa. Sistemas sintetizadores de fala podem ler 'em voz alta' um texto escrito, estes podem ser utilizados em interfaces adaptadas para deficientes visuais e também em serviços automatizados de informação por telefone.

Pesquisa em reconhecimento e síntese da fala do português brasileiro tem sido realizada, no Brasil, através do projeto Spoltech *Advancing Human Language Technology in Brazil and the United States Through Collaborative Research on Spoken Language Systems,* (http://www.ucs.tche.br/lpv/spoltech/) coordenado pelo Prof. Dante Barone da Universidade Federal do Rio Grande do Sul.

4.2. Corretores ortográficos e gramaticais

As últimas versões de editores de texto (*Microsoft Word*, por exemplo) possuem um subsistema de correção ortográfica e gramatical que verifica se cada uma das palavras digitadas pertence ao vocabulário da linguagem e verifica algumas construções gramaticais das frases como, por exemplo, as regras de concordância da língua. Esses sistemas trabalham com um léxico que pode ser estendido pelo usuário, e a correção

[5] http://visl.hum.sdu.dk/visl/

gramatical aponta erros relativos ao uso da crase, de colocação pronominal, concordância verbal, pontuação, uso de prefixos etc.

A versão do corretor ortográfico da língua portuguesa, presente hoje no *Microsoft Word,* foi desenvolvida com apoio da Itautec/Philco no Núcleo Interinstitucional de Lingüística Computacional (NILC-USP), através do projeto *ReGra,* coordenado pela Profª Maria das Graças Volpe Nunes [NUN 00]. Esse sistema de correção gramatical, além de possuir um módulo gramatical que realiza a análise sintática, é baseado em um conjunto de regras heurísticas que servem para detectar, por exemplo, os erros de uso de crase. O sistema também possui um outro módulo, chamado de módulo mecânico, que trata erros de fácil detecção, tais como: palavras e símbolos de pontuação repetidos, presença de símbolos de pontuação isolados, uso não balanceado de parênteses e aspas, capitalização inadequada como início de frase com letra minúscula, e ausência de pontuação no final da sentença.

4.3. Tradutores automáticos

Há diversos sistemas tradutores que se tornaram produtos comerciais (*Translator Pro*, *Tradunet*), ou que são de distribuição gratuita e disponíveis pela Internet (*Alta Vista*, *Intertran*, *GO Translator*, *Enterprise Translator Server*). Esses sistemas de tradução são considerados preliminares, no sentido de que fazem uma tradução não refinada; é freqüente a ocorrência de erros e imperfeições no resultado final obtido. Uma análise detalhada da qualidade do resultado obtido por esses tradutores é apresentada em [OLI 00]. Diferentes metodologias podem ser empregadas na tradução automática, entre elas, podemos citar os sistemas diretos, os sistemas transferenciais e os sistemas *interlingua*. Os sistemas diretos buscam correspondências diretas entre as palavras, enquanto os sistemas de transferência efetuam a análise sintática da frase da língua de origem e, através de regras de transferência sintática, constróem a representação sintática na língua alvo. Os sistemas interlinguais trabalham com uma representação intermediária entre as línguas origem e alvo que, em princípio, pode ser utilizada na tradução de quaisquer línguas. Mais informações sobre tradutores automáticos podem ser obtidos em [JUR 00].

4.4. Geradores de textos e resumo

A geração de textos pode ser vista como o processo inverso da interpretação: o gerador recebe como entrada elementos de conteúdo e objetivos de comunicação, para produzir um texto lingüisticamente correto. Deve determinar o que será dito e de que forma, organizando o discurso e as frases. Um dos desafios da área é o processo de planejamento envolvido na geração do discurso. Questões relacionadas ao planejamento podem ser abordadas com o aporte das teorias envolvendo agentes [BEA 91]. Em [BAR 96] uma introdução à área de geração de linguagem natural é apresentada.

Os geradores de resumo constituem um recurso bastante útil no processo de busca de informação. Resumos gerados automaticamente podem auxiliar uma pessoa na decisão sobre a relevância de um determinado documento. Diferentemente da geração de textos, a geração de resumos deve proporcionar o máximo de informação no mínimo de espaço, e isso envolve o estudo do uso da linguagem para veicular informação de forma concisa. Nesse tipo de aplicação, dá-se uma relação interessante com técnicas

estatísticas, através da identificação dos modos como as palavras são utilizadas pela análise de grandes corpora.

4.5. Interfaces em linguagem natural

Uma das aplicações mais comuns para interface em linguagem natural é a manipulação de base de dados, onde um sistema de processamento de linguagem natural serve de intermediário entre o usuário e a base de dados, traduzindo as instruções apresentadas em linguagem natural para a linguagem específica do sistema de gerenciamento de dados. Tais interfaces podem ser baseadas na linguagem escrita ou falada e são, usualmente, denominadas ‘sistemas de perguntas e respostas’. Sistemas de perguntas e respostas eficientes são geralmente relativos a um domínio de aplicação bem especificado e limitado, muitas vezes delimitando-se a interação a palavras-chaves. Exemplos que podem ser dados aqui são informações sobre viagens de uma determinada estação ferroviária, e serviço bancário.

4.6. Recuperação de informação

A recuperação de informação é a área de aplicação envolvida com a obtenção de documentos relevantes dado um determinado tema, e não está diretamente envolvida com a obtenção de uma informação específica ou com a obtenção de resposta a uma dada pergunta. Recuperação de informação pode, então, ser definida como sendo o conjunto de técnicas que servem ao propósito de encontrar documentos relevantes de acordo com uma necessidade de informação. Em geral, essas técnicas são constituídas por indexação, busca, filtragem, organização, tratamento de múltiplas línguas e também múltiplas mídias. Existem duas abordagens principais distintas, a busca por metadados (cabeçalhos ou palavras-chaves que descrevem o conteúdo dos documentos) ou por conteúdo. Metadados podem ser adicionados aos documentos manualmente (o que é dispendioso e muito subjetivo) ou automaticamente (onde se obtém uma qualidade razoável, mas não muito alta). Abordagens baseadas em conteúdo atingem, em geral, melhores resultados. Note, no entanto, que são abordagens baseadas em técnicas estatísticas que medem a similaridade de textos e da consulta, e não em compreensão de texto. A compreensão automática de texto é ainda uma área com baixa efetividade em domínios irrestritos. Pode excepcionalmente ser uma opção mais adequada em domínios restritos. Uma obra importante que apresenta bons elementos para os estudos da área é [BAE 99].

4.7. Extração de informação

Enquanto sistemas de recuperação de informação encarregam-se de encontrar documentos relevantes em relação a um determinado tema, sistemas de extração de informação encarregam-se de analisar e transformar a maneira de apresentação da informação contida em um conjunto de documentos relevantes, isolando informações relevantes contidas em determinados segmentos, e apresentando a informação encontrada em um formato coerente. Sistemas de extração de informação podem ‘ler’ um texto não estruturado e coletar informação a ser armazenada em um banco de dados tradicional.

Extração de informação é uma área de interesse para pesquisas em lingüística computacional, pois possui tarefas e problemas bem definidos. Os sistemas utilizam

textos reais e a performance dos sistemas pode ser avaliada de acordo com a performance humana na execução da mesma tarefa. Tais sistemas motivam, dessa maneira, os pesquisadores em lingüística computacional a migrarem, de sistemas de pequena escala e dados artificiais, para sistemas de larga escala e dados lingüísticos reais. A área de extração de informação popularizou-se com a série de competições americana intitulada *Message Understanding Conferences* (MUCs). Mais informação sobre a área pode ser obtida em [COW 96].

4.8. Avaliação de sistemas de processamento de linguagem natural

Algumas das aplicações discutidas nesta seção apresentam uma tradição maior em avaliação de resultados produzidos pelos sistemas, notoriamente a recuperação de informação é uma delas. Sistemas de recuperação são usualmente avaliados em termos de alcance e precisão (ou *recall* e *precision*). O alcance, nesse contexto, mede o número de documentos relevantes encontrados para uma consulta, entre o conjunto total de documentos relevantes (documentos relevantes encontrados / total de documentos relevantes existentes) e a precisão mede o número de documentos realmente relevantes entre os indicados como relevantes pelo sistema (documentos relevantes encontrados / documentos encontrados). Sistemas de extração de informação também têm sido sistematicamente avaliados, e conferências têm sido organizadas em forma de competição para a apresentação desses sistemas (*Message Understanding Conference* MUC-3 1991, MUC-4 1992, MUC-5 1993, MUC-6 1994). Diferentes aplicações podem desenvolver ou utilizar critérios próprios. Os critérios considerados podem ter cunho lingüístico, operacional ou econômico. Uma avaliação de desempenho de tradutores automáticos para a tradução de Inglês-Português-Inglês [OLI 00], por exemplo, faz uma avaliação lingüística considerando os níveis: léxico, sintático e semântico-pragmático. Em [NUN 00] uma avaliação de desempenho para o corretor ortográfico da língua portuguesa (ReGra) é apresentada.

Corpora anotados são um recurso importante no processo de avaliação de sistemas, uma nova técnica proposta pode ser avaliada de acordo com um corpus anotado em nível morfológico, sintático ou semântico. Certas informações lingüísticas relacionadas a um discurso podem ter um caráter mais subjetivo, o caso da co-referência é um exemplo, dificultando a tarefa de anotação de corpus e, conseqüentemente, a de avaliação de sistemas. Nesse caso, algumas medidas têm sido empregadas para avaliar o grau de concordância entre diferentes sujeitos realizando a anotação de um corpus, de acordo com um dado esquema. Um sistema, nesse caso, pode ser avaliado com uma anotação derivada de várias anotações, ou o desempenho pode ser medido através do grau de concordância entre sistema e anotação manual. Em [POE 98] uma avaliação de desempenho de um sistema de resolução de co-referência, com essas características, é apresentada.

4.9. Processamento de linguagem natural e sistemas multi-agentes

Uma abordagem computacional alternativa, para os sistemas de processamento da linguagem natural, é a organização em sociedades de agentes. Essa abordagem multiagentes foi estudada, para a língua portuguesa, através do projeto NALAMAS [SIL 98, STR 99], desenvolvido em cooperação por cinco grupos brasileiros e um grupo português. No estudo realizado, foi dada ênfase a diferentes fenômenos lingüísticos, entre os quais ambigüidade, anáforas e elipses, e sua resolução através de

uma abordagem multiagentes. Foram também desenvolvidos, utilizando uma plataforma adequada, protótipos de solução multi-agentes para esses fenômenos em português.

A respeito desse esforço, algumas conclusões são interessantes de mencionar. Primeiramente, é necessária uma migração de todos os analisadores e demais ferramentas disponíveis, de modo a orientá-los a uma concepção em agentes, e de modo a projetar adequadamente os conhecimentos coletivos e individuais dos agentes. Só então é possível passar-se à proposta de soluções mais específicas.

Pode-se considerar que, nas situações em que é necessária a articulação entre múltiplas fontes de conhecimento, como é o caso da resolução de anáforas, na interpretação, ou o processo de planejamento, na geração de linguagem, a abordagem multiagentes se mostra promissora. Porém, a aplicabilidade dessa abordagem a fenômenos específicos não significa que ela seja interessante à totalidade dos níveis de análise.

5. Processamento de corpus

O trabalho realizado na área de lingüística de corpus reúne, compila e organiza repositórios de trechos de linguagem escrita ou falada, naturalmente e espontaneamente gerados e que servem de base para a pesquisa lingüística. Este trabalho, só foi tornado possível com a ajuda do computador e, portanto, data dos inícios dos anos 60.

Recentemente novos repositórios têm sido criados de maneira que informação lingüística sobre os dados seja adicionada ao corpus. A prática de adicionar informação lingüística interpretativa a um corpus eletrônico, contendo dados lingüísticos de fala ou escrita, é chamada de anotação de corpus. Um caso típico e familiar de anotação de corpus é a etiquetagem gramatical (comumente conhecida como *part-of-speech tagging*). Nesse caso, uma etiqueta é associada a cada palavra do corpus, indicando sua classe gramatical. Assim como estão divididos os níveis de estudo lingüísticos e os diferentes problemas abordados em lingüística computacional, a anotação de corpus também se divide em anotação morfológica ou gramatical, anotação sintática, semântica, e de discurso. Trabalhos nessa área podem estar relacionados à construção manual de corpus anotado, a criação de padrões para a anotação de corpus, criação de ferramentas para auxílio à marcação manual de corpus, criação de ferramentas para marcação automática ou semi-automática de corpus. Esta última envolvendo criação de sistemas que façam a interpretação lingüística de textos, em nível morfológico, sintático ou de discurso, dependendo do tipo de marcação a ser realizada, utilizando muitas vezes um corpus marcado com um tipo de informação, para a realização da marcação de um novo tipo de informação lingüística. Uma ilustração dos diferentes tipos de anotação de corpus é dada a seguir, para pequenos trechos de um discurso.

5.1. Anotação de corpus

5.1.1. Anotação gramatical

O exemplo que segue apresenta a etiquetagem morfológica (ou *POS tagging*), que associa a cada palavra de um texto uma etiqueta contendo sua classe gramatical e sua forma lexical canônica.

ela **_PPR_ele** sofre **_V_sofrer** grande **_ADJ_grande** rejeição **_N_rejeição**

de **_PREP_de** os **_ART_o** governadores **_N_governador**

Este formalismo foi utilizado em projeto desenvolvido pelo Grupo de Língua Natural[6] do Centro de Investigação em Inteligência Artificial (CENTRIA) da Universidade Nova de Lisboa em Portugal, coordenado pelo Prof. Gabriel Pereira Lopes. No exemplo dado acima foi utilizado esquema de códigos para anotação morfológica, que inclui os seguintes códigos:

ADJ	ADJetivo
ART	ARTigo
N	Nome (substantivos comuns)
PR	Pronome Relativo
PREP	Preposição
V	Verbo

5.1.2. Anotação sintática A seguir, é apresentada, para o mesmo trecho visto acima, uma análise sintática de acordo com a gramática de restrições (*Constraint Grammar*), tal como utilizado pelo projeto VISL *Visual Interactive Syntax Learning.*

=SUBJ:pron-pers(F 3S NOM/PIV) ela

=MV:v-fin(PR 3S IND) sofre

=ACC:np

==>N:adj(M/F S) grande

==H:n(F S) rejeição

==N<:pp

===H:prp(<sam->) de

===P<:np

====>N:art(<-sam> <artd> M P) os

====H:n(M P) governadores

Para esclarecer o exemplo dado acima, listamos a seguir algumas dessas convenções, extraídas das páginas do projeto[7].

SYNTACTIC TAGS (etiquetas sintáticas)

SUBJ	subject (sujeito)
ACC	accusative (direct) object (objeto direto acusativo)
MV	main verb (verbo principal)

[6] http://pc-gpl.di.fct.unl.pt/~glint

[7] O conjunto completo de símbolos utilizado para a marcação de análise sintática desse projeto é apresentado em http://visl.hum.sdu.dk/visl/pt/portsymbol.html.

N<	postnominal adject (attaches to the nearest NP-head to the left, that is not an adnominal itself) (adjunto pós nominal)
P<	argument of preposition (argumento da preposição)
H	head (núcleo)

5.1.3. Anotação sintática parcial (sintagmas nominais)

O exemplo apresentado aqui mostra uma anotação de corpus parcial, correspondendo ao conjunto de sintagmas nominais extraídos do trecho: *ela sofre grande rejeição de os governadores.*

['**SN',**['**N',**ela]].

['**SN',**os,['**N',**governadores]].

['**SN',**grande,['**N',**rejeição],['**SP',**de,['**SN',**os,['**N',**governadores]]]] .

As marcas utilizadas são SN para indicar sintagma nominal, N para indicar núcleo do sintagma, e SP para indicar sintagma preposicional. Esta anotação parcial e notação são utilizadas pelo projeto ANACORT – Anotação automática de co-referência textual, em desenvolvimento na Universidade do Vale do Rio dos Sinos e coordenado pela Profª Renata Vieira. Uma descrição da construção do corpus do projeto ANACORT com anotação parcial de sintagmas nominais é apresentada em [VIE 00].

5.1.4. Anotação de discurso

O exemplo a seguir ilustra a marcação de co-referência, ou seja, indicação de expressões em um discurso que se referem ao mesmo objeto ou entidade.

> São remotas as chances de aprovação < **coref:de ID = "de_01"** > da atual proposta de projeto de reforma tributária **</coref:de >.** Embora esteja ainda em fase de discussão, < **coref:de ID = "de_02"** > ela **</coref:de >** sofre grande rejeição dos governadores.
>
> **<coref: link type = "ident" href = "coref.xml#id(de_02)">**
>
> **<coref: anchor href = "coref .xml#id(de_01)">**
>
> **</coref link>**

O formalismo apresentado acima segue as diretrizes apresentadas pelo projeto MATE[8] - *Multilevel annotation tools engineering* - para a marcação de co-referência [POE 00]. O esquema de anotação proposto por esse projeto foi desenvolvido com base na linguagem de marcação XML, onde "coref" (*coreference*) indica um elemento ou relação de co-referência no discurso, "de" (*discourse entity*) indica uma entidade de discurso, "link", uma ligação entre um elemento e um antecedente identificado por "anchor". O projeto MATE tem por objetivo desenvolver ferramentas e um padrão para anotação de corpora de diálogos falados.

A anotação de corpus apresenta múltiplas funcionalidades, muitas das vantagens de se ter acesso a tais recursos lingüísticos são revertidas para a pesquisa e

[8] http://mate.mip.ou.dk/

desenvolvimento da área de lingüística computacional. Um corpus marcado com informação sobre a classe gramatical pode ser útil, por exemplo, a um sistema de síntese de fala, onde a diferenciação entre a categoria substantivo ou verbo pode indicar uma alteração na pronúncia (*o jogo, eu jogo,* por exemplo). Outras aplicações (extração de informações lexicográficas, tradução automática, ou recuperação de informação) podem também se beneficiar de tais recursos. Em [GAR 97] uma apresentação completa da área de anotação de corpus é dada.

Cabe ainda mencionar, como exemplo de trabalho realizado nessa área, o projeto *TychoBrahe Parsed Corpus of Historical Portuguese*, desenvolvido na UNICAMP e USP [BRI 99].

6. Conclusão

A área de lingüística computacional envolve um grande conjunto de atividades voltadas ao objetivo de tornar possível a comunicação com as máquinas utilizando as habilidades naturais de comunicação humana. A pesquisa na área inclui o reconhecimento, interpretação, tradução e geração de linguagem e requer um esforço de convergência entre várias disciplinas: lingüística, computação e psicologia, por exemplo. A área tem um papel muito importante para a sociedade de informação. Avanços no processamento de fala, texto e imagem são necessários para tornar mais acessível, e possibilitar o melhor uso, da grande quantidade de informação que está hoje disponível na rede mundial de computadores. É uma área promissora, especialmente em relação à língua portuguesa. É importante considerar a necessidade de formação de recursos humanos nessa área relativamente nova, que atualmente, no Brasil, se faz presente mais em cursos de pós-graduação do que na graduação.

7. Bibliografia

[ALL 00] ALLAN, J. Natural Language Processing for Information Retrieval. **Tutorial of the NAACL/ANLP Language Technology Joint Conference** in Seattle, Washington, April 29, 2000.

[ALL 95] ALLEN, J. **Natural Language Understanding**. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc., 1995. 654p.

[AUS 62] AUSTIN, J.L. **How to do things with words**. Oxford, Clarendon Press, 1962.

[BAE 99] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.

[BAR 96] BARROS, F. e ROBIN, J. Processamento de linguagem natural. Jornada de Atualização em Informática JAI, **Anais do XVI Congresso da Sociedade Brasileira de Computação** 1996.

[BEA 91] BEARDON, C. et al. **Natural Language and Computational Linguistics**. Melksham-Wiltshire, England, Ellis Horwood Ltd., 1991.

[BOU 98] BOUILLON, P. **Le traitement automatique des langues**. Bruxelles, Duculot, 1998. 245p.

[BRI 95] BRILL, E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. **Computational Linguistics**, 21(4), 543-566. 1995.

[BRI 99] BRITTO, H. & FINGER, M. "Constructing a parsed corpus of Historical Portuguese". ACH/ALLC-99 **International Humanities Computing Conference**. University of Virginia, Charlottesville, Virginia. Junho, 1999.

[CHO 56] CHOMSKY, N. Three models for the description of language. **IRE Transactions** PGIT, 2. (pp. 113-124), 1956.

[CHO 57] CHOMSKY, N. **Syntactic structures**. The Hague, Mounton. 1957.

[COO 83] COOPER, R**. Quantification and Syntactic Theory**. Reidel, Dordrecht. 1983

[COV 94] COVINGTON, M. A. **Natural language processing for Prolog programmers.** New Jersey, Prentice Hall. 1994.

[COW 96] COWIE, J. and LEHNERT, W. Information Extraction. **Communications of the ACM**, Vol.39, Nº 1, January, 1996.

[DOW 81] DOWTY, D.R., WALL, R.E. and PETERS, S. **Introduction to Montague semantics**. Dordrecht, D. Reidel Pub. Co. 1981.

[FUC 92] FUCHS, C., LE GOFFIC, P. **Les Linguistiques Contemporaines**. Paris, Hachette, 1992. 158p.

[GAR 97] GARSIDE, R., LEECH, G. and McENERY, A. (Eds.) **Corpus annotation**: linguistic information from computer text corpora. Longman, London, 1997.

[GAS 00] GASPERIN, C. V. **Fundamentos do Processamento Estatístico da Linguagem Natural**. Porto Alegre, PPGCC-PUCRS: Trabalho Individual, 2000.

[GAZ 82] GAZDAR, G. Phrase Structure Grammar. In Jacobson and Pullum, (eds): **The Nature of Syntactic Representation**. Reidel, Dordrecht. 1982.

[GAZ 85] GAZDAR, G., KLEIN, E. PULLUM, G. and SAG, I. **Generalized Phrase Structure Grammar**. Basil Blackwell, 1985.

[GEA 52] GEACH, P. and BLACK, M. **Translations from the philosophical writings of Gottlob Frege**. Totowa, Barnes & Noble Books. 1952.

[GON 00] GONZALEZ M. **O léxico gerativo de Pustejovsky sob o enfoque da recuperação de informações**. Porto Alegre, PPGCC-PUCRS: Trabalho Individual. 2000.

[GON 00b] GONZALEZ, M. A. I. **Representação Semântica de Sentenças em Linguagem Natural e sua aplicação na Recuperação de Informações**. Porto Alegre, PPGCC-PUCRSR: Trabalho Individual, 2000.

[GRE 96] GREEN, D. W. et al. **Cognitive Science: an introduction.** Cambridge, Blackwell Publishers Ltd., 1996.

[GRI 68] GRICE, H. P. Utterer's meaning, sentence meaning, and word-meaning. **Foundations of Language**, 4, (pp. 1-18). 1968

[GRI 75] GRICE, H. P. Logic and conversation. In: Cole, P. and Morgan, J.L. (Eds.) **Syntax and semantics**, Vol. 3: Speech acts (pp. 225-242). New York, Academic Press, 1975.

[JUR 00] JURAFSKY, D., MARTIN, J. **Speech and Language Processing**. New Jersey, Prentice-Hall, 2000. 934p.

[KAY 79] KAY, M. Functional grammar. In **Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society**, 1979.

[KAM 93] KAMP, H. and REYLE, U. **From discourse to logic**. Dordrecht, Kluwer.

[KLA 96] KLAVANS, J. L. **The balancing act : combining symbolic and statistical approaches to language**. Cambridge: MIT Press, 1996.

[KOW 93] KOWALTOWSKI, T., LUCCHESI, C. L. **Applications of finite automata representing large vocabularies**. Software Practice and Experience, 23(1), 15-30, 1993.

[LEW 96] LEWIS, D. D. and SPARCK JONES, K. Natural language processing for information retrieval. **Communications of the ACM**, Vol.39, Nº 1, January, 1996.

[MAN 99] MANNING, C. and SCHÜTZE, H. **Foundations of Statistical natural language processing.** Cambridge, MA: The MIT Press, 1999. 680p.

[NIJ 88] NIJHOLT, Anton. **Computers and languages – theory and practice**. Amsterdam: Elsevier, 1988. 482p.

[NUN 99] NUNES, M. G. V. et al. Introdução ao Processamento das Línguas Naturais. **Notas didáticas do ICMC Nº 38**, São Carlos, 88p., 1999.

[NUN 99] NUNES, M. G. V. e OLIVEIRA, N. O. O processo de desenvolvimento do revisor gramatical ReGra. SEMISH **Anais do XX Congresso da Sociedade Brasileira de Computação**, Curitiba, 2000.

[OLI 00] OLIVEIRA, N. O., et al. A critical analysis of the performance of English-Portuguese-English MT systems. **Anais do V Encontro para o Processamento do Português Escrito e Falado**. (pp. 85-92) Atibaia-SP, Novembro, 2000.

[PET 99] PETRY, T. O., STRUBE DE LIMA, V. Considerando o uso de 'centering' na resolução de referências anafóricas pronominais em português. In: **Actas do IV Encontro para o Processamento da Língua Portuguesa Escrita e Falada** (PROPOR'99), Évora – Portugal. 1999.

[POE 00] POESIO, M. Coreference. **MATE Dialogue Annotation Guidelines-Deliverable D2.1**, January 2000. (http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html).

[POE 98] POESIO, M. and VIEIRA, R. A Corpus-based investigation of definite description use. In **Computational Linguistics**, Vol. 24 (2): 183-216. 1998.

[POL 94] POLLARD, C. and SAG, I. A. **Head-driven phrase structure grammar.** Chicago, The University of Chicago Press. 1994.

[PUS 95] PUSTEJOVSKY, J. **The generative lexicon**. Cambridge, MIT Press. 1995.

[SAT 01] SANT'ANNA V. M. e STRUBE DE LIMA V. L. Cálculo de Referências Anafóricas Pronominais Demonstrativas na Língua Portuguesa Escrita. In: **Anais do Encontro Nacional de Inteligência Artificial** (ENIA). Fortaleza, 30 jul a 3 ago, 2001.

[SEA 69] SEARLE, J. R. **Speech acts: an essay in the philosophy of language**. Cambridge, Cambridge University Press. 1969.

[SHA 49] SHANNON, C. E. and WEAVER, W. **The mathematical theory of communication**, Illinois, University of Illinois Press. 1949.

[SHI 71] SHIEBER, S. M. The design of a computer language for linguistic information. **Proceedings of the 10th International Conference on Computational Linguistics** COLING, (pp. 362-366), California USA, 1984.

[SIL 98] SILVA, J.L.T., ABRAHÃO, P.R.C., STRUBE DE LIMA, V. Integrating morphological, syntactical and semantical aspects through multi-agent cooperation. In: F. Oliveira (ed.). **Advances in Artificial Intelligence: 14th Brazilian Symposium on Artificial Intelligence** - SBIA'98, Porto Alegre, Brazil, November 4-6, Proceedings. Lecture Notes in Artificial Intelligence 1515. pp. 83-92. Springer-Verlag. ISBN 3-540-65190-X 1998.

[STR 99] STRUBE DE LIMA, V. et al. 1999. NALAMAS – Natural Language Multi-Agent Systems: studying the subject through NALAMAS project. In: V.

Almeida et al. (eds.). In: **Proceedings of the PROTEM-CC'99 Projects Evaluation Workshop**, Rio de Janeiro, Brazil, May 05-07. pp. 73-98.

[TRA 99] TRASK, R. L. **Key concepts in Language and Linguistics**. Routledge, London, 378p., 1999.

[VER 97] VERHAREN, E. M. **A language-action perspective on the design of cooperative information agents**. Proefschrift Katholieke Universiteit Brabant Tilburg, Nederlands. PhD Thesis.

[VIE 98] VIEIRA, R. **Definite description processing in unrestricted text**. PhD Thesis. Division of Informatics, Edinburgh University. Edinburgh, UK.

[VIE 00] VIEIRA, R. Extração de sintagmas nominais para processamento de co-referência. **Anais do V Encontro para o Processamento do Português Escrito e Falado**. (pp. 165-174) Atibaia-SP, Novembro, 2000.

[VIL 95] VILLAVICENCIO, A. **Avaliando um rotulador estatístico de categorias morfo-sintáticas para a língua portuguesa**. Porto Alegre, CPGCC-UFRGS, 1995. Dissertação de Mestrado.

[WOO 70] WOODS, W. A. Transition network grammars for natural language analysis. **Communications of the ACM**, 13(10), (pp. 591-606), 1970.