DataGramaZero - Revista de Ciência da Informação - v.3 n.1 fev/02 *ARTIGO 05*

A Ambiqüidade na Recuperação da Informação

Ambiguity in information retrieval

por Marisa Bräscher

Resumo: Discutem-se questões relativas à extração de informações contidas em textos completos e posterior recuperação, por meio de métodos de tratamento automático da linguagem natural. Além da extração de palavras do texto, procura-se manter as relações de significado que estas palavras possuem no contexto do discurso em que ocorrem. Assim, defende-se o tratamento de determinados fenômenos lingüísticos que afetam a qualidade da recuperação, como o da ambigüidade. Como referencial teórico-metodológico para efetuar a análise e organização sintático-semântica de conteúdos, utilizam-se a Gramática de Valências de Borba e a teoria de Gráficos Conceituais de Sowa. Emprega-se um sistema de tratamento automático da linguagem natural – o Zstation – em um corpus constituído de documentos oficiais do Mercosul, para testes de desambiguação. Conclui-se que um sistema de recuperação da informação em linguagem natural pode solucionar determinados tipos de ambigüidades quando dispõe de informações relativas à valência sintático-semântica das unidades lexicais que compõem um enunciado. Os resultados obtidos demonstram ser possível introduzir procedimentos automáticos de solução de ambigüidades em sistemas de tratamento da linguagem natural.
Palavras-chave: Recuperação da Informação; Tratamento Automático da Linguagem Natural; Ambigüidade; Valência Sintático-Semântica; Gráficos Conceituais

Abstract: TIssues relative to information extraction from complete texts and subsequent retrieval by means of automatic natural language treatment methods are discussed. Besides extracting words from the text, the relationship of significance that these words have in the context of the speech in which they occur is attempted to be preserved. An information retrieval system using natural language should be able to treat given linguistic phenomena that affect the quality of information, such as, for instance, the issues of ambiguity. The Valence Grammar and the Conceptual Graphics are used as theoretical and methodological. An automatic natural language treatment system – Zstation – is utilized, as well as a pool of official documents concerning the Mercosul, for the ambiguity solutions tests. The conclusion is reached that a natural language treatment system can solve certain types of ambiguities when information is available regarding the syntactic-semantic valence of the lexical units that compose an enunciation. The results obtained show that it is possible to introduce automatic procedures for solving ambiguities in a natural language treatment system.
Keywords: Information Retrieval; Document Analysis; Natural Language Processing; Ambiguity; Valence Grammar; Conceptual Graphs

INTRODUÇÃO

As tecnologias da informação vêm provocando mudanças profundas nos processos tradicionais de comunicação científica, quase que eliminando o espaço de tempo entre a produção e a disseminação dos textos científicos. Essas mudanças, conseqüentemente, afetam os processos de tratamento da informação utilizados pela Ciência da Informação. Observa-se hoje uma tendência ao desenvolvimento de mecanismos que possibilitam a disponibilização dos documentos no momento de sua produção, em muitos casos pelo próprio autor. Como exemplo desta tendência podem ser citados a Biblioteca Digital de Teses e Dissertações Eletrônicas, da Virginia Tech [i] e os Arquivos Abertos [ii]. Ambos fornecem aos autores padrões e ferramentas para produção e submissão eletrônica de documentos, possibilitando a disseminação imediata das informações disponibilizadas nestes repositórios.

Apesar das mudanças ocorridas nos processos de produção, tratamento e disseminação de informação, alguns problemas enfrentados pelos sistemas tradicionais de recuperação da informação continuam presentes nas ferramentas de busca atuais e ganham maior amplitude e complexidade. Como ressalta Chen [iii] isto de deve a diferentes fatores: variações nas estruturas e formatos de bases de dados,

diferentes formas de documentos disponibilizados (texto, audio e vídeo) e abundância de conteúdos multilíngües nas aplicações da Web. Acrescente-se, ainda, a estes aspectos, a multidisciplinaridade dos conteúdos disseminados na rede.

Considerando estes fatores e o contexto atual de produção e disseminação eletrônica de documentos, as pesquisas realizadas na área de recuperação da informação concentram-se, de maneira geral, no desenvolvimento de ferramentas que possibilitem a extração do conteúdo diretamente dos textos completos dos documentos disponibilizados eletronicamente. No entanto, ferramentas de busca que utilizam palavras como pontos de acesso ao conteúdo têm se mostrado ineficientes, fato este observado pela quantidade de informação irrelevante recuperada por motores de busca da Web.

Assim, os trabalhos mais recentes na área baseiam-se na premisssa de que ferramentas de busca, ao fazerem uso da linguagem natural, necessitam de conhecimento sobre o significado das expressões que são tratadas e das relações que se estabelecem entre elas. Essas ferramentas devem, ainda, ser capazes de tratar determinados fenômenos lingüísticos que afetam a qualidade da recuperação, como o da ambigüidade, a qual é tratada no âmbito deste trabalho.

WEB SEMÂNTICA

A necessidade de recuperação de informações armazenadas em grandes repositórios de informação disponíveis na Internet e de responder com maior precisão às buscas realizadas diretamente pelos usuários finais, têm levado a um esforço no sentido de adicionar informação semântica às páginas Web. Procura-se, desta forma, como afirma Cranefield [iv], aumentar a eficiência e a seletividade dos motores de busca e de outros tipos de ferramentas de processamento automático de documentos.

As propostas de incorporação de informação semântica em sistemas de busca aplicam abordagens distintas, enfatizando um ou outro aspecto da análise lingüística e utilizando diferentes métodos de organização de bases de conhecimento [cf. v]. Doerr [vi] e Hunter [vii] defende o uso de tesauros, que organizam termos e associam conceitos em redes semânticas, como uma ferramenta importante para a busca de informação eletrônica, ressaltam, no entanto, a necessidade de tratar problemas relativos à interoperabilidade semântica entre diferentes tesauros e a necessidade de desenvolvimento de metavocabulários (*metadata vocabularies*) para permitir o intercâmbio e a busca de informação em diferentes aplicações e domínios.

Nas pesquisas realizadas no âmbito do projeto Digital Libraries Iniciative (DLI) [iii] procura-se recuperar os avanços em diversas áreas, tais como reconhecimento, segmentação e indexação de objetos; análise semântica em sistemas de tratamento automático da linguagem natural; representação do conhecimento e interação homem-máquina, tendo como principal objetivo tornar possível a interoperabilidade semântica nas bibliotecas digitais.

A questão da interoperabilidade semântica torna-se importante no âmbito das pesquisas relacionadas à busca na Internet, uma vez que os diferentes repositórios de informação eletrônica (bibliotecas digitais, bases de dados, etc.) utilizam sistemas próprios de organização semântica das informações. O desafio que se coloca neste sentido, como afirmam Berners-Lee et al [viii] é fornecer uma linguagem que expresse dados e regras para raciocínio sobre esses dados de forma que as regras de qualquer sistema de representação do conhecimento possam ser exportadas para a Web.

Esta é a proposta da Web Semântica, que visa fornecer uma estrutura de conteúdo significativo para as

páginas Web, criando um ambiente onde os *softwares agents* possam realizar tarefas sofisticadas para os usuários. A Web Semântica utiliza-se da flexibilidade da estrutura RDF (Resource Description Framework), na qual é possível descrever o conteúdo da informação disseminada na rede, fazendo-se afirmações sobre determinado objeto e identificando suas propriedades e valores. Cada objeto ou assunto é identificado por um Identificador Universal de Registro (URI) que assegura que as palavras na Web estejam relacionadas a apenas uma definição. [viii]

A Web semântica utiliza-se ainda das ontologias para possibilitar a recuperação de conceitos. Uma ontologia na Web Semântica possui uma taxonomia e um conjunto de regras de inferência. A taxonomia define as classes de objetos e as relações que se estabelecem entre eles. Forma-se assim uma estrutura onde propriedades são atribuídas a determinadas classes e os objetos que pertencem a esta classe herdam suas características.
A solução de ambigüidades e a obtenção de maior precisão na recuperação de informações disponíveis na Web constitui-se numa das principais preocupações dos estudos da Web Semântica. Berners-Lee et al [viii] afirmam que a ambigüidade pode ser solucionada atribuindo-se diferentes URIs para cada conceito de uma palavra. Assim, os motores de busca poderão encontrar páginas que se refiram a conceitos específicos e não todas as páginas nas quais a palavra ambígua é utilizada. Outros tipos de ambigüidades, no entanto, podem ocorrer no conteúdo de documentos disponíveis na Web, interferindo também na precisão da recuperação da informação.

AMBIGÜIDADE

Entende-se ambigüidade como uma expressão da língua (palavra ou frase) que possui vários significados distintos, podendo, conseqüentemente, ser compreendida de diferentes maneiras por um receptor. [ix; x] A ambigüidade ocorre quando palavras ou frases podem gerar mais de uma interpretação de seu significado, como nos seguintes exemplos:

> Ex.1: na frase *O arquivo está precisando de manutenção*, a ambigüidade latente da palavra arquivo induz à interpretação de um arquivo como móvel, um arquivo como conjunto de documentos ou de um arquivo como instituição.
>
> Ex. 2 - na fraseologia *Neutralização de contaminação com leite* , a ambigüidade permite interpretar que a neutralização é feita com leite ou que a contaminação é causada pelo leite.

A ambigüidade causa ruído na recuperação da informação, pois, sob um mesmo termo, o usuário encontrará informação relevante e irrelevante. No exemplo 1, o usuário recuperará informação sobre manutenção de arquivo em três direções semânticas distintas: conjunto de documentos, instituição e móvel. Qual desses significados respondem à sua pergunta? No exemplo 2, a ambigüidade sintática não permite, num sistema de recuperação, decidir entre os assuntos *neutralização de contaminação* e *neutralização com leite*.

Ao encontrar diferentes significados possíveis de serem extraídos de uma frase ou palavra, o sistema de recuperação necessita distinguir um destes significados, determinando, segundo o contexto, qual o significado a ser aplicado, obtendo, dessa maneira, maior precisão na resposta dada ao usuário.

A ambigüidade pode ser ocasionada por diversos fatores[1]: polissemia, homografia, policategorização,

relação contextual e estrutura sintática das frases. Segundo o fator que a ocasiona, a ambigüidade pode ser classificada em diferentes tipos. Pela sistematicidade e clareza com que distingue os tipos de ambigüidades, adota-se, neste estudo, a classificação de Fuchs [x], sintetizada a seguir.

I) Ambigüidade morfológica: ocorre quando não é possível classificar determinada forma quanto à categoria gramatical. Este tipo de ambigüidade é ocasionado pela policategorização – em que palavras pertencem a mais de uma categoria gramatical, como proposta , que pode ser ou substantivo, ou adjetivo ou verbo.

II) Ambigüidade lexical: ocorre quando há mais de uma interpretação possível do significado de uma unidade lexical. Este tipo de ambigüidade é provocado por :

homografia : ocorre por meio da « *colisão acidental entre as formas de dois signos lingüísticos distintos* ». [x; p.9]. Ex.: cobre (metal) ; cobre (do verbo cobrir)

polissemia : ocorre quando uma só e mesma expressão envolve significados distintos, sendo um único signo lingüístico; é a própria expressão que é ambígua, à medida que possui uma forma à qual corresponde uma pluralidade de significados. [x]. Ex.: arquivo (móvel, instituição, conjunto de documentos).

III) Ambigüidade sintática : ocorre na estruturação da frase em constituintes hierarquizados, quando se definem as ligações que se estabelecem entre os sintagmas. As frases preposicionais são uma das fontes mais freqüentes de ambigüidade sintática. Alguns exemplos ilustram este tipo de ambigüidade:

Ex. 3: *Eu li a notícia sobre a greve na universidade*. (ou eu li a notícia e eu estava na universidade, ou a greve ocorre na universidade)

Ex. 4: *A professora de dança espanhola*. (ou a professora é espanhola, ou a dança é espanhola)

IV) Ambigüidade predicativa : ocorre na interpretação das relações temáticas que articulam predicado, argumentos e participantes. Exemplos :

Ex. 5: *A crítica deste autor*. (autor = ou objeto da crítica, ou agente da crítica)

Ex. 6: *Eu a deixei feliz*. (feliz = ou atributo do sujeito ou atributo do objeto)

V) Ambigüidade semântica : ocorre quando há mais de uma interpretação possível para o relacionamento dos termos na frase, como, por exemplo, no cálculo dos operadores de negação e de quantificação :

Ex. 7: *Ela não chora mais porque ele partiu*. (ou ela chorava porque ele havia partido, ou ela parou de chorar uma vez que ele já foi embora)

Ex. 8: *Um rio corre através de cada país europeu*. (ou um único rio corre através de todos os países, ou diferentes rios correm através de diferentes países)

VI) Ambigüidade pragmática : relaciona-se ao cálculo dos valores enunciativos, à reconstrução destes valores, que estão ligados à situação do falante no momento da enunciação, como por exemplo :

Ex. 9: *Os pássaros voam*. (referência geral ou específica?)

Ex. 10: *Paulo vai à escola*. (ele é estudante ou ele está indo à escola neste momento?)

Como demonstram esses exemplos, a ambigüidade pode ser ocasionada por diferentes fenômenos lingüísticos situados nos níveis morfológico, lexical, sintático, semântico e pragmático. A solução destes problemas depende do objetivo de um sistema de recuperação da informação e das bases de conhecimento disponíveis neste sistema.

DESAMBIGUAÇÃO NA RECUPERAÇÃO DA INFORMAÇÃO

Denomina-se desambiguação[2] o processo pelo qual uma ambigüidade é solucionada. Este processo exige diferentes níveis de conhecimentos lingüísticos e extralingüísticos.

A ambigüidade morfológica, causada por policategorização, por exemplo, pode ser solucionada pela análise do co-texto imediato que circunda a palavra policategorial, recorrendo-se apenas a conhecimento morfossintático (categoria gramatical, concordância e combinações sintáticas entre constituintes da frase, entre outros). Na frase *O governo aumentou o imposto*, governo e imposto, por estarem precedidos do determinante, são interpretados corretamente pelo sistema como substantivos e não como verbos (formas flexionadas dos verbos governar e impor).

Certos casos de polissemia são solucionados por meio de conhecimento semântico. Ao dispor das informações:

* comprar é uma ação que exige objeto comercializável ;
* móveis são objetos que podem ser comprados ; e
* arquivo é um tipo de móvel ;

um sistema recuperação em linguagem natural pode atribuir corretamente o significado móvel a arquivo na frase *Maria comprou um arquivo para seu escritório*[3]

Algumas ambigüidades predicativas são solucionadas pela introdução de traços semânticos que restringem os papéis temáticos desempenhados pelos argumentos de um predicado. Como exemplifica Borba [xi], o sintagma nominal *A observação da criança* é ambíguo, mas *A observação do quadro* não, uma vez que, pelo traço -humano, *quadro* não estabelece uma relação agente de observar.

Existem, portanto, determinados tipos de ambigüidades que podem ser solucionadas automaticamente, pois os conhecimentos necessários para desambiguá-las são passíveis de modelização aplicando-se métodos de tratamento automático da linguagem natural.

A solução de ambigüidades em sistemas de recuperação em linguagem natural tem por objetivo determinar que escolhas são mais adequadas considerando-se o contexto onde ocorre a ambigüidade. Como afirma Fuchs [x], toda forma à qual podem ser associados vários significados é virtualmente ambígua (ambigüidade virtual) quando considerada isoladamente, fora de todo contexto de uso.

Quando esta forma é analisada num contexto, ela pode se tornar unívoca, ou pode ser considerada efetivamente ambígua (ambigüidade efetiva).

Sistemas desenvolvidos para desambiguar aplicam diferentes técnicas de tratamento automático da linguagem natural e aplicam regras formais segundo a abordagem lingüística e o modelo de representação do conhecimento adotados pelo sistema. A complexidade das regras utilizadas varia em função do tipo de ambigüidade que se visa solucionar.

O processo de desambiguação automática é mais complexo que o de solução de ambigüidades realizado por um receptor humano. O recurso ao contexto em sistemas de recuperação de informação em linguagem natural é restrito, uma vez que o contexto constitui-se no conjunto de conhecimentos que o sistema possui num determinado momento da análise. Nem todo tipo de informação contextual pode ser representado formalmente e, portanto, nem todo tipo de ambigüidade pode ser resolvido nesses sistemas.

A pesquisa relatada neste artigo trata de diferentes tipos de ambigüidades e propõe a desambigüação por meio de tratamento sintático-semântico, utilizando gráficos conceituais como estrutura de representação de conhecimento.

GRÁFICOS CONCEITUAIS COMO MODELO DE REPRESENTAÇÃO DE CONHECIMENTO

A teoria dos gráficos conceituais (GCs) começou a ser desenvolvida por Sowa em 1968, quando escreveu um trabalho de final de curso para Minsky. Neste trabalho, Sowa aplicou a idéia de fluxogramas para criar um modelo de representação de conhecimento em Inteligência Artificial que se utiliza de caixas e círculos para gerar Gráficos Conceituais (GCs). Na década de 70, Sowa inicia um trabalho de pesquisa sobre gráficos conceituais como linguagem de representação do conhecimento no *Systems Research Institute* da IBM. O resultado deste trabalho é publicado, em 1976, no *IBM Journal of Research and Development*. Oito anos após, Sowa [xii] publica seu livro *Conceptual Structures*, apresentando a teoria de GCs como hoje é conhecida. [xiii]

Como modelo de representação do conhecimento que utiliza uma notação em gráficos, os GCs são, para Sowa [xii; p. 7] « *uma síntese dos gráficos existenciais de Peirce, dos gráficos de dependência de Tesnière e das redes semânticas da Inteligência Artificial.* » Os GCs formam uma linguagem de representação do conhecimento e são constituídos por gráficos que possuem dois tipos de nós :

> a) os **conceitos**, representados por retângulos ou por colchetes [CONCEITO], correspondem a conteúdos de pensamento ; representam entidades, ações ou estados que possam ser descritos em termos de linguagem; e
>
> b) as **relações**, representadas por círculos com uma flecha de entrada e outra de saída ou entre parênteses => (RELAÇÃO) =>, simbolizam as ligações existentes entre os conceitos e demonstram os papéis que cada entidade desenrola.

Para Sowa [xii; p.20], « *os gráficos conceituais formam uma base semântica da linguagem natural e representam modelos do mundo real ou de um mundo possível.* » No esquema da figura 1,

demonstra-se como funciona o mecanismo implícito no triângulo do conceito, com os GCs servindo de ligação entre o referente e o significante, onde:

a) as regras de sintaxe mapeam gráficos para sentenças em LN e mapeam sentenças para gráficos.

b) os arcos dos gráficos correspondem à função da palavra e a casos relacionais da LN. No exemplo da figura 1, EST e LOC são, respectivamente símbolos das relações ‘estado’ e ‘local’

c) os nós dos gráficos são conceitos intensionais de indivíduos que devem existir no mundo real ou em algum mundo hipotético.

O mesmo gráfico, gerado a partir de um processo de percepção, serve de representação para as frases expressas nas duas línguas - português e francês. Os GCs constituem-se, portanto, numa linguagem universal e independente, no nível da estrutura profunda.

CONCEITOS NOS GCS

Nos gráficos conceituais, um conceito é um objeto que possui um tipo e um referente que especifica exatamente que espécie do tipo precedente o conceito representa.
O tipo do conceito não é necessariamente muito distante (do ponto de vista semântico) do conceito representado. Por exemplo, o tipo do conceito gato é GATO[4] e não MAMÍFERO, apesar de gato ter como hiperônimo mamífero. Esta relação de hiperonímia encontra-se representada por uma rede, chamada *treillis* de conceitos, na qual é estabelecida a hierarquia entre tipos. A relação representada nessa hierarquia é uma relação de ordem de grandeza que se estabelece entre tipos de conceitos e não entre conceitos individuais.

Existem, assim, diversas famílias de conceitos, isto é, conjuntos de conceitos que têm o mesmo hiperônimo. Esses conceitos são ditos do mesmo tipo. Tomando-se como exemplo o *tipo* FRUTA, pode-se dizer que laranja, pêra e banana são do tipo FRUTA - fruta é um hiperônimo de laranja, pêra e banana.

A hierarquia de tipos é um ordenamento parcial definido a partir de um conjunto de etiquetas de tipo. O símbolo <= determina a ordem hierárquica. Os termos subtipo e supertipo são utilizados para designar a posição dos conceitos na hierarquia, como abaixo :

Se X < Y, então :
X é um subtipo de Y, e
Y é um supertipo de X.
Se X <= Y e X < = Z, então :
X é um subtipo comum de Y e Z.
Se X >= Y e X >= Z, então :
X é um supertipo comum de Y e Z.

Na hierarquia de tipos, assim como em outras estruturas hierárquicas gênero/espécie baseadas em Aristóteles, os subtipos herdam as propriedades de seus supertipos. Um *treillis* de conceitos deve ter

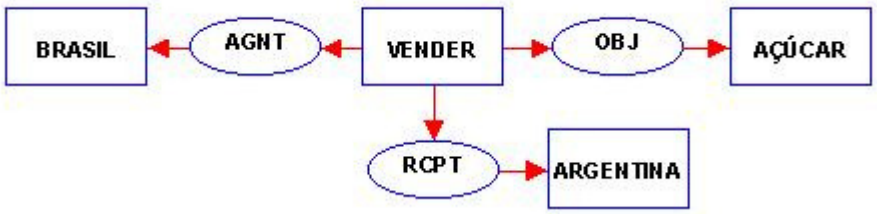
supertipos e subtipos comuns. Para indicar os tipos de conceitos de forma linear, utiliza-se a seguinte notação :

[<tipo> :<referente>]
Ex. : [INSTITUIÇÃO : 'Embratel']

RELAÇÕES NOS GCS

As relações conceituais definem o papel de cada conceito num GC. São as ligações que se estabelecem entre os conceitos do gráfico. Podem ter um número qualquer de arcos, sendo que a relação mais comum é *díade*[5].

A representação em diagramas não é fácil de ser construída quando se estabelecem várias relações entre os conceitos do GC. Dessa forma, Sowa [xii] propõe uma notação linear que substitui os diagramas, escolhendo como « cabeça » o conceito ao qual se ligam maior número de arcos. O gráfico conceitual



é representado linearmente assim :

[VENDER]-
(AGNT) => [BRASIL]
(OBJ) => [AÇUCAR]
(RCPT) => [ARGENTINA]

Os GCs devem ser lidos de acordo com o sentido das flechas. No exemplo dado, lê-se : VENDER tem por agente BRASIL, por objeto AÇÚCAR e por receptor ARGENTINA. Esse tipo de representação segue uma sintaxe própria, como o emprego do hífen após a caixa do conceito VENDER no exemplo acima, para indicar que as relações que se estabelecem com este conceito estão listadas nas linhas subseqüentes. A sintaxe completa utilizada nessa notação é descrita por Sowa [xii], no apêndice A6 de seu livro. As duas notações - a linear e a em gráfico - são exatamente equivalentes e podem ser traduzidas automaticamente para outras formas de lógica ou de representação do conhecimento.

Para evitar falsas combinações entre conceitos e relações num gráfico conceitual, Sowa [xii] introduziu o conceito de gráfico canônico. Diz-se que um gráfico é canônico quando « r*epresenta situações reais ou possíveis num mundo externo.* » [xii; p.91] A construção de um conjunto coerente de GCs que formam uma base de conhecimento é feita a partir dos Gráficos Conceituais Canônicos (GCCs) que exprimem as restrições semânticas do domínio representado.

Os gráficos canônicos são utilizados num analisador semântico para orientar a escolha de certas combinações entre relações e conceitos. Como afirma Sowa [xii; p.222], eles « *fornecem preferências semânticas para certas combinações e reforçam restrições que bloqueiam outras combinações.* » Esse tipo de orientação auxilia na solução de casos de ambigüidade sintática, porque as restrições semânticas levam à escolha da interpretação correta da frase.

BASES DE CONHECIMENTO PARA TRATAMENTO SINTÁTICO-SEMÂNTICO DE AMBIGÜIDADES

Neste artigo, descreve-se, de forma resumida, pesquisa realizada por Bräscher [xiv], na qual se utiliza conhecimentos sintático-semânticos organizados com base na gramática de valências de Borba [xi] para solução de ambigüidades em textos de língua portuguesa. Estes conhecimentos constituem-se, basicamente de :

a) conhecimento sintático : características morfossintáticas dos elementos que representam, na estrutura superficial, uma relação predicado/argumento; função sintática destes elementos e como eles organizam-se sintaticamente;

b) conhecimento semântico : características dos conceitos (traços semânticos); relações semânticas (hiperonímia, sinonímia, p.ex.) e relações temáticas (agente, ação, objeto, entre outras).

Essas informações sintático-semânticas encontram-se armazenadas em bases de conhecimento de acordo com o formalismo adotado no sistema Zstation [xv]. O Zstation constitui-se num sistema de tratamento automático da linguagem natural, cuja idéia básica é que, para desempenhar uma tarefa, como analisar uma sentença, faz-se necessário coletar toda informação sobre esta sentença, quanto a propriedades semânticas e morfológicas das palavras, possíveis grupos de palavras e frases, e conexões possíveis entre eles, até que o conhecimento coletado permita propor uma ou várias interpretações.

Os módulos especialistas deste sistema são responsáveis por tarefas específicas. Cada módulo tem acesso a uma base de conhecimento em forma declarativa. Os módulos especialistas são os seguintes:

I) Geração morfossintática
Constrói formas corretas a partir de **lemas**[6] de acordo com variáveis morfossintáticas (número, tempo, etc.), extraídas de uma gramática morfológica que descreve como as formas são geradas. O programa de geração morfossintática procura primeiramente a qual modelo morfológico - protótipo de palavra - um lema morfológico é associado. Depois ele procura pela gramática de geração associada a esse modelo e aplica a gramática ao lema, gerando as diversas formas possíveis. Duas fontes de conhecimento são necessárias: a gramática morfológica e a base de dados que associa lemas a modelos, ambas são programadas usando formalismo declarativo.

II) Análise morfossintática
Encontra o lema morfológico correspondente para cada forma no texto, e sua categoria morfossintática (substantivo, verbo, pronome, adjetivo, Tc). Sua tarefa reduz-se a consultar uma base de dados que contém todos os lemas do dicionário de lemas.

III) Análise sintagmática
Extrai todos os tipos de grupos necessários para a análise sintática da sentença ou de unidades de texto maiores. Há uma diferença importante entre análise sintagmática e análise sintática. O programa de análise sintagmática basicamente extrai tipos específicos de grupos (grupo nominal,

preposicional, verbal, adverbial, etc.). Na análise sintática, o objetivo é identificar as ligações entre grupos ou frases, definindo os papéis destes grupos na frase: sujeito, objeto1, objeto2, etc.

IV) Análise semântica

Procura, previamente, todos os conceitos que podem ser associados a um lema morfológico, para, então, obter as informações semânticas necessárias à análise semântica. Num segundo estágio, o módulo determina todas as restrições semânticas que são associadas a determinado conceito. Os parâmetros semânticos são definidos sob forma de traços individuais e de traços de classe e são estruturados em redes semânticas. Nestas redes os conceitos constituem-se em nós aos quais podem ser ligados atributos semânticos e outros conceitos hierarquicamente relacionados.

O cálculo das ligações entre grupos é um processo complexo para o qual tanto a informação sintagmática quanto a semântica são requeridas. Os conhecimentos lingüísticos relativos à análise sintática são formulados de maneira a considerar o conjunto de parâmetros sintáticos e semânticos que podem ser atribuídos a um lema específico. Dessa maneira, a cada lema morfológico podem ser associados um ou vários conceitos.

Para efetuar cada tipo de análise, o Zstation utiliza diferentes tipos de ferramentas lingüísticas que são definidas e construídas pelo usuário do sistema. Essas ferramentas são baseadas em formalismo de ampla aplicação de maneira que é possível construir dicionários e gramáticas para diferentes línguas, como francês, italiano, português, espanhol, inglês e alemão.

DICIONÁRIO AUTOMÁTICO

Um dicionário no Zstation é constituído de um conjunto de lemas e de dados lingüísticos referentes a eles, como ilustrado no exemplo a seguir :

brasileiro
{CPT=brasileiro0
MOD=amigo
VSM=
VGR=
APD=
}
{CPT=brasileiro0
MOD=belo
VSM=
VGR=
APD=$qual
  arg(0,rel=CHRC,cat=adj,fonct=modSub,conds=[ ])
}

Para cada registro são previstos, no dicionário de base, os seguintes dados lingüísticos :

a) Identificador do conceito (CPT) : conjunto de caracteres que simbolizam o conceito representado pelo lema. No formalismo adotado, o conceito é representado adicionando-se o símbolo 0 ao final da cadeia de caracteres escolhida para representar o conceito. O CPT possibilita a localização de um conceito numa Ontologia e é utilizado em qualquer análise automática efetuada pelo sistema que aplique o conceito como variável. No exemplo dado, *brasileiro0* representa o conceito do lema *brasileiro*.

b) Modelo morfológico (MOD) : lema escolhido para representar uma classe de lemas que, pertencendo a uma mesma categoria, sofre a mesma flexão quanto ao tempo, ao modo e à pessoa, para verbos, e quanto ao gênero e ao número para as demais categorias. No exemplo, *amigo* é o modelo morfológico do lema *brasileiro* na condição de substantivo e *belo* na condição de adjetivo.

c) Argumentos (APD) : contêm parâmetros sintático-semânticos relacionados ao lema de entrada. Constituem-se numa série de enunciados que estabelecem condições sintático-semânticas a serem observadas no momento da análise. Os argumentos são definidos com base na valência sintática e semântica do lema. Cada argumento é estruturado da seguinte maneira :

(Code, rel=R,cat=C,fonct= F,conds=[r(...)], em que:

Code = código de prioridade que pode ser 0 para um argumento facultativo ;1 para argumento obrigatório representado em termos de relação conceitual ; 2 para argumento obrigatório que não é passível de representação em relação conceitual e 3 para argumento proibido (regra de bloqueio).
rel = relação temática estabelecida com o conceito do lema na Ontologia.
cat = categoria morfossintática do argumento.
fonct = função sintática do argumento.
conds = condições de validação intralingüísticas, que não podem ser deduzidas da Ontologia utilizada. São definidas em forma de relações conceituais.

Um lema pode ter um ou vários blocos de dados, de acordo com as categorias gramaticais às quais pertence. Os blocos são delimitados por colchetes. O lema *brasileiro* possui dois blocos de dados, um para cada uma das categorias gramaticais às quais pertence. O primeiro bloco registra o substantivo, atribuindo ao lema o modelo morfológico *amigo*, o segundo indica que o lema pode também ser um adjetivo, para o qual se aplica o modelo *belo*.

Os dados descritos em a, b e c, foram utilizados nos dicionários construídos no curso da pesquisa realizada. Além destes, podem ser registrados nos dicionários do Zstation : variáveis semânticas intralingüísticas (VSM) e variáveis gramaticais intralingüísticas (VGR), que são variáveis próprias à determinada língua que está sendo tratada. Esses dados são utilizados sobretudo em pesquisas multilíngües.

No exemplo de entrada do lema ação, ilustra-se uma entrada completa no dicionário. Este lema é monocategorial porque é sempre um substantivo, portanto, seu modelo morfológico é o mesmo em todos os blocos de dados. Porém, por ser polissêmico, são-lhes atribuídos vários conceitos. Para cada conceito existem restrições sintático-semânticas que são definidas nos diferentes argumentos.

ação
{CPT=praticaração0
MOD=ação
VSM=
VGR=
APD= arg(0,rel=FIN,cat=sub_de,fonct=modN,conds=[ ])
}
{CPT=titcred0
MOD=ação
VSM=
VGR=
APD= arg(0,rel=ORIG,cat=sub_de,fonct=modN,conds=[ ])
}
{CPT=convpojur0
MOD=ação
VSM=
VGR=
APD= arg(0,rel=AGNT,cat=sub_de,fonct=Spsagt,conds=[ ])
arg(0,rel=OBJ,cat=sub_contra,fonct=Spcomp1,conds=[ ])

O argumento atribuído à *atividade0* indica que um conceito representado por substantivo precedido da preposição de (sub_de) possui uma relação finalidade com o conceito de *atividade0* do lema ação. Este substantivo é um modificador de N (N é o lema de entrada), pois indica uma característica de N. Em *titcred0* (título de crédito), o conceito expresso pelo 'sub_de' indica a origem (ORIG) do título e constitui-se num modificador do nome ação. No sentido de convocar poder jurisdicional (*convpojur0*), ação, possui outra estrutura argumental :

a) um agente (AGNT) representado por um sub_de, que está em relação subjetiva com o predicado (fonct=Spsagt) ;

b) um objeto (OBJ) do ato de convocar, indicado por sub_contra, que se constitui no primeiro e único complemento.

O argumento de *convpojur0* informa também que um substantivo precedido da preposição para (cat=sub_para) indica com que finalidade (rel=FIN) convoca-se o poder jurisdicional. Este substantivo funciona como modificador, não sendo parte da matriz valencial.

Nos argumentos, as condições sintático-semântica são enunciadas. O detalhamento dos parâmetros sintáticos é feito na Gramática de Variáveis, e dos parâmetros semânticos, na Ontologia.

GRAMÁTICA MORFOLÓGICA

As gramáticas morfológicas no Zstation reúnem o conjunto de lemas selecionados como modelos morfológicos para os demais lemas incluídos num dicionário de base. Cada entrada de uma gramática inclui : o modelo morfológico, a categoria gramatical, as variáveis (pessoa e tempo para verbos e

gênero e número para demais categorias aos quais se aplicam) e a regra morfológica a ser aplicada.

O modelo *amigo* exemplifica uma entrada da Gramática Morfológica Portuguesa criada no âmbito da pesquisa:

amigo
CAT=sub VARS=[masc,sing] REGS=[ ]
CAT=sub VARS=[masc,plur] REGS=[+s]
CAT=sub VARS=[fem,sing] REGS=[-o,+a]
CAT=sub VARS=[fem,plur] REGS=[-o,+as]

Na gramática morfológica, CAT identifica a categoria gramatical do modelo, VARS as variáveis morfológicas e REGS a regra a ser aplicada segundo a variação definida. No exemplo, o lema 'amigo' constitui o modelo morfológico de todos os substantivos que formam o masculino/plural com acréscimo do 's' ; o feminino com a substituição do 'o' pelo 'a', e o feminino/plural com a troca do 'o' pelo 'as'.

A aplicação automática do modelo morfológico adequado a cada lema do dicionário permite que outros programas do Zstation identifiquem, nos textos que estão sendo analisados automaticamente, todas as formas possíveis de determinado lema. A utilização do modelo morfológico reduz o número de entradas de um dicionário automático. Faz-se necessária apenas uma entrada para cada lema, as demais formas são geradas e reconhecidas automaticamente.

GRAMÁTICA DE ARGUMENTOS

Esta gramática especifica como se efetuam as ligações entre os constituintes relacionados a determinada função sintática. As regras são enunciadas segundo a sintaxe do Zstation, como descrito a seguir:

r(X,Cat,Fonct,F,Ops), em que:
X = forma a ser encontrada
Cat = categoria associada à X
Fonct = função associada à X
F = forma de referência
Ops = operações lingüísticas

As regras da gramática de argumentos possibilitam que o sistema identifique e analise, nos enunciados do corpus, as seqüências que devem ser interpretadas segundo os parâmetros estabelecidos nos argumentos. A interpretação dos enunciados recorre também às informações semânticas descritas na Ontologia.

As regras estabelecidas para o argumento *convpojur0* do exemplo 2, demonstram o uso da gramática :

Regra 1 : r(X,sub_de,Spsagt,F[match(F,de,X)])

Regra2 : r(X,sub_contra,Spcomp1,F[match(F,contra,X)])

Na primeira parte da regra, que está fora do parênteses, encontram-se as variáveis a serem interpretadas. As informações incluídas nos parênteses orientam o sistema a interpretar as variáveis estabelecidas.

A regra 1, por exemplo, determina que, encontrando uma seqüência F + de + substantivo, o sistema deve interpretar de + *substantivo* como sintagma preposicional em relação de sujeito agente (Spsagt). Da mesma maneira será interpretada a regra 2: ao encontrar as seqüências indicadas pelo comando *'match'*, o sistema deverá interpretá-las como sintagma preposicional em relação de complemento (Spcomp1). A interpretação dos enunciados recorre também às informações semânticas descritas na Ontologia.

ONTOLOGIA

A relação temática definida num argumento é especificada na Ontologia, que se constitui numa representação linear dos gráficos conceituais. A Ontologia representa objetos e relações de um domínio específico. Cada conceito é uma entrada na Ontologia, sendo acompanhado dos tipos de relações que podem ser com ele estabelecidas. As relações podem indicar uma propriedade do conceito (relação ISA) ou as relações que este possui com outros conceitos ou classes de conceitos.

Os conceitos definidos para o lema ação, no dicionário, foram registrados da seguinte maneira na Ontologia :

praticação0
r(0,isa,+abstrato0)
r(0,FIN,+ações0)

titcred0
r(0,isa,+produto comerciável0)
r(0,ORIG,+instifin0)
r(0,POSS,+animado0)

convpojur0
r(0,ISA,ação-processo)
r(0,AGNT,+animado0)
r(0,OBJ,+animado0)

A cada relação podem ser especificadas, se necessário, as características ou categorias conceituais que delimitam os tipos de conceitos com os quais esta relação pode ser estabelecida. Cada relação contém três tipos de informação :

a) um código que indica se uma declaração é obrigatória (1) ou facultativa (0) ;

b) um identificador de relação temática; e

c) um conceito, característica ou classe de conceito com o qual se estabelece a relação temática. O símbolo **+** indica os que são aceitos, aqueles com os quais a relação pode ser estabelecida, e o símbolo **-** informa os que não são aceitos.

Os conceitos indicados nas relações devem ser também incluídos na Ontologia, até se chegar às classes mais genéricas da cadeia hierárquica, cujo supertipo é U. A Ontologia forma um *"treillis"* de conceitos estabelecendo-se, portanto, um mecanismo de hereditariedade. Os subtipos herdam as propriedades de seus supertipos. Indicando-se na Ontologia que *banco* é uma *instfin0* (instituição financeira), este conceito será aceito para a relação ORIG do conceito *titcred0*.

O conceito *instfin0* é um subtipo de *instituição0*, que, por sua vez, é um subtipo de *entidades animadas*. Segundo esta cadeia hierárquica, qualquer conceito do tipo *instituição0* é aceito na relação AGNT de *convpojur0*.

Na pesquisa realizada, as características eleitas para se estabelecer a Ontologia levaram em conta as áreas de assunto do Mercosul, tema do corpus de pesquisa.

DESAMBIGUAÇÃO APLICANDO TRATAMENTO SINTÁTICO-SEMÂNTICO

O conjunto de dados registrados no Dicionário, na Gramática Morfológica, na Gramática de Argumentos e na Ontologia foram utilizados para efetuar-se o tratamento sintático-semântico de enunciados do corpus de pesquisa, verificando a ocorrência de ambigüidades e se estas foram solucionadas ou não pelo sistema Zstation.

Fornecendo o enunciado : *A empresa vende produtos ao consumidor*, o sistema gera o seguinte Gráfico Conceitual:

[VENDER]-
(AGNT) => [EMPRESA]
(OBJ) => [PRODUTOS]
(BEN) => [CONSUMIDOR]

Aplicando as regras de formação de gráficos conceituais, o Zstation é capaz de analisar, também, os seguintes enunciados :

Venda de gás ao consumidor.
O Brasil venderá café ao Paraguai.
O exportador vendeu vinho à loja.

Como *gás, café e vinho* são tipos de produtos, são aceitos como argumento da relação objeto (OBJ) de vender; *Brasil* e *exportador* são aceitos como agentes (AGNT) de vender e *consumidor*, *Paraguai* e *loja* pelo traço +animado, são também aceitos como argumentos da relação beneficiário (BEN) de vender. O sistema analisará corretamente estes enunciados pois:

a) dispõe da informação, no Dicionário de Formas, de que venderá e vendeu são formas do lema vender, e as reconhece como verbos;

b) reconhece também as categorias gramaticais das demais formas do enunciado, uma vez que estas se encontram indicadas pelos modelos morfológicos informados para cada lema no Dicionário de Base;

c) interpreta os papéis temáticos e as funções sintáticas desempenhadas por cada palavra que compõe os enunciados, com base nos argumentos indicados no Dicionário; nas estruturas sintáticas descritas na Gramática de Argumentos e nas relações e nos traços semânticos informados na Ontologia.

A análise em GCs realizada pelo sistema permite testar e avaliar se as restrições sintático-semânticas, registradas nas bases de conhecimento, são suficientes para solucionar casos de ambigüidades ocasionados por homografias e polissemias.

As seqüências constituídas de nome abstrato de ação[7] + sintagmas preposicionais (Sprep), extraídas automaticamente do corpus pelo sistema Zstation, são utilizadas como massa de teste de solução de ambigüidades por meio de tratamento sintático-semântico.

Pelos resultados da análise em GCs, é possível verificar se ocorre ou não ambigüidade. Em caso afirmativo, realiza-se o teste de desambigüação, aplicando restrições sintático-semânticas. Dessa forma, é possível concluir se o sistema é capaz de selecionar um significado entre as alternativas de interpretação possíveis.

Os gráficos canônicos ‘a’ e ‘b’ informam, respectivamente, que vender exige como objeto (OBJ) um produto comerciável e que ação é um objeto comerciável, um tipo de ação ou um efeito. O nome ação, portanto, é polissêmico, mas, dispondo das informações contidas nos gráficos, o Zstation é capaz de escolher o significado ‘valor financeiro’ para esta forma, no enunciado “*regras de preferência para os casos de venda de **ações** e aumento do capital social*”.

a) [VENDER]-
(OBJ) => [PRODUTO COMERCIÁVEL]
b) [AÇÃO]-
(ISA) => [PRODUTO COMERCIÁVEL]
(ISA) => [AGIR]
(ISA) => [EFEITO]

Os testes realizados demonstraram que outros tipos de ambigüidades também podem ser solucionadas por meio de tratamento sintático-semântico, como exemplificado a seguir.

Ambiguidade predicativa

Este tipo de ambigüidade ocorre quando mais de um tipo de relação temática pode ser estabelecido entre predicado e argumentos. Os nomes abstratos de ação que indicam ação-processo e que possuem os argumentos objeto e agente introduzidos pela preposição de podem apresentar ambigüidade

predicativa.

Quando o argumento não possui o traço +animado, não ocorre ambigüidade, sendo corretamente interpretado como objeto, uma vez que um inanimado não pode funcionar como agente. Resta, portanto, apenas uma interpretação, como em *aprovação da tarifa, avaliação da proposta* e *regulamentação das normas*.

No entanto, quando o argumento possui o traço +animado, ocorre dupla interpretação. O argumento pode ser interpretado como agente ou objeto, como em *administração dos estados, aprovação da comissão, designação do diretor* e *regulamentação do estado*. Nestes casos, duas análises em GCs são possíveis, como no exemplo de designar, em que a primeira interpretação equivale a 'o diretor designa alguém' e a segunda a 'alguém designa o diretor'.

DESIGNAR-
(AGNT) - [DIRETOR]
(OBJ) - [+ANIMADO]
DESIGNAR
(AGNT) - [+ANIMADO]
(OBJ) - [+DIRETOR]

Segundo Borba [xi], este tipo de ambigüidade ocorre porque o sintagma preposicional em relação subjetiva pode se tornar contíguo ao nome abstrato, passando a ser introduzido por de, por causa do apagamento[8] do sintagma preposicional em relação objetiva e, ainda, devido à possibilidade de apagamento do sintagma preposicional em relação subjetiva.

Quando não há apagamento, não ocorre ambigüidade predicativa, como em *aprovação pelo Organismo Executor do relatório final*.

APROVAR
(AGNT) - [ORGANISMO EXECUTOR]
(OBJ) - [RELATÓRIO FINAL]

Com base nos testes realizados em ocorências do corpus de pesquisa, é possível afirmar que a ambigüidade predicativa pode ser solucionada, por meio de tratamento sintático-semântico, quando outros elementos do contexto oferecem restrições que orientem a escolha da interpretação correta.

Polissemia

A polissemia dos verbos subjacentes também ocasiona polissemia nos nomes abstratos de ação correspondentes, como no caso do verbo determinar, que pode significar *'estabelecer'*, *'fixar'* e *'ordenar'*.

Pela análise da valência, observa-se que a natureza dos argumentos pode, em certos contextos, solucionar a ambigüidade do nome *determinação*, como nos exemplos a seguir:

a) traço semântico +princípios permite atribuir o significado 'estabelecer' em *determinação de requisitos*;

b) traço semântico +valor permite atribuir o significado 'fixar' em *determinação do montante*; e

c) traço semântico +ação permite atribuir o significado 'ordenar' em *determinação da cessação*;

Estes traços são considerados pelo sistema no momento da análise em GCs, permitindo a interpretação correta do significado do nome abstrato de ação.

Ocorre também polissemia entre o conceito do verbo subjacente e o de uma entidade concreta ou abstrata envolvida na ação como o agente, o objeto, o resultado ou o instrumento utilizado na ação. Os nomes *notificação* (ação ou resultado de notificar?), *pedido* (ação ou resultado de pedir?) e *administração* (ação ou agente de administrar?), exemplificam este tipo de polissemia.

Nestes casos, o traço semântico do argumento ou a valência de outro elemento do contexto permitem a desambigüação, como nos exemplos a seguir:

a) *A Presidência Pro Tempore da Comissão remeterá aos demais Estados-Parte cópia das notificações referidas no art...*

O argumento objeto (OBJ) de copiar é representado, na estrutura superficial do nome abstrato de ação, por um substantivo precedido da preposição de (sub_de). Este argumento é preenchido por um conceito do tipo documento. Com base nestes dados, o sistema decidiu corretamente pela interpretação 'documento' (resultado da ação) para o nome notificação.

b) *Os resultados da investigação deverão ser comunicados às autoridades do país importador em um prazo não superior a quarenta e cinco (45) dias corrigidos, contados a partir da data de recebimento do pedido.*

Como a ação de receber exige um argumento objeto com o traço +concreto, o sistema decidiu corretamente pelo significado 'documento', para o nome pedido, descartando o significado 'ação de pedir'.

c) *As informações fornecidas à administração aduaneira ou por esta obtida...*

Os argumentos do conceito *fornecer0* permitiram a solução da polissemia de *administração*, no contexto exemplificado acima. A preposição *a* introduz o constituinte que preenche o argumento beneficiário. Este argumento exige um conceito com o traço +animado. Dessa maneira, o significado 'instituição' foi o escolhido pelo sistema.

Nos casos exemplificados, assim como em outros analisados, o sistema pôde escolher uma única interpretação. Isto foi possível pois a valência sintático-semântica dos constituintes que precederam ou

sucederam os nomes forneceram parâmetros sintático-semânticos que restringiram o tipo de conceito que pode preencher seus argumentos.

Em outros contextos, porém, os constituintes não ofereceram elementos para a desambiguação, como nos exemplos d, e, f:

> d) *As mercadorias somente poderão ser descarregadas ou transportadas mediante autorização da autoridade aduaneira...*
>
> e) *inutilizar ou dificultar a operação de equipamento...*
>
> f) *Para suas comunicações oficiais, a Secretaria disporá de facilidades não menos favoráveis que as outorgadas pela República às missões diplomáticas...*

No exemplo *d*, as duas interpretações são possíveis : +documento e +ação. No entanto, a relação que ocorre entre a autorização e autoridade aduaneira permanece sendo a mesma : a autoridade aduaneira é quem concedeu a autorização.

No exemplo *e* , a ambigüidade é causada pela ocorrência de polissemia nos dois nomes - *operação* e *equipamento*. O nome operação, no sentido de 'manobrar' exige um argumento com o traço +equipamento e, no sentido de 'efetuar operação', um argumento com o traço +ação. Neste caso, as duas interpretações são possíveis, uma vez que a forma equipamento admite os dois traços.

O nome *comunicação*, no exemplo *f*, admite tanto a interpretação dinâmica (ação da Secretaria se comunicar) quanto a estática (documentos do tipo 'comunicação' elaborados pela Secretaria). A solução deste tipo de ambigüidade só é possível quando outros elementos do contexto oferecem parâmetros sintático-semânticos que favorecem uma ou outra interpretação. Na frase « *Não serão objeto de censura a correspondência e outras comunicações oficiais da Secretaria.* », o nome *censura* favorece a interpretação *documento* para *comunicações*.

Homografia

Quando não ocorre relação semântica entre as entidades e as ações representadas pelo nome abstrato de ação, estes foram classificados como homógrafos, como no caso do nome *ação*, que pode significar 'título de crédito' (ex.: venda de ações); 'praticar ação' (ex.: ação de articulação); 'convocar o poder jurisdicional' (ex.: ação administrativa ou judicial) ou 'efeito' (ex.: substância de ação hormonal)

Nos testes efetuados, o nome ação foi desambiguado pelo sistema nas seguintes situações :

a) quando o argumento do próprio nome introduzido pela preposição *de* possuía o traço +ações[9], como em ações de apoio ; ações de articulação. Nestes casos o conceito 'praticar ação' foi selecionado pelo sistema ;

b) quando este nome se constituía num argumento ou especificador de outro elemento predicador, como em :

i) *venda de ações* ; *transferência de ações*. Os nomes *venda* e *transferência* admitem como argumento apenas o conceito de título de crédito ;

ii) *os países-membros realizarão ações necessárias*. Entre os diferentes conceitos de *ação*, o verbo *realizar* admite o conceito 'praticar ação' ;

iii) *substância de ação hormonal*. Como característica do nome *substância*, apenas o conceito 'efeito' é aceito para o nome *ação*.

Não foi possível solucionar a ambigüidade quando o argumento de ação possuía o traço + animado, como em *ações do estado*. Neste caso, o sistema admitiu as seguintes interpretações :

[TÍTULO DE CRÉDITO]-
(POSS) - [ESTADO]
[TÍTULO DE CRÉDITO]-
(ORIG) - [ESTADO]
[PRATICAR AÇÃO]
(AGNT) - [ESTADO]
[CONVOCAR PODER JURISDICIONAL]
(AGNT) - [ESTADO]

Para todos os tipos de ambigüidades identificados nos testes realizados, a desambiguação foi possível quando a valência sintática e semântica do nome ou de outros constituintes presentes no contexto, bem como os traços semânticos de seus argumentos, forneceram restrições que orientaram a escolha de uma entre as possíveis interpretações.

CONCLUSÃO

Os avanços tecnológicos influenciam a área de informação e conduzem ao surgimento de novas técnicas de representação e recuperação de conteúdo. No contexto tecnológico atual, há tendência para o desenvolvimento de sistemas inteligentes de recuperação de informação com base em processamento de linguagem natural, em função da disponibilidade de textos completos em máquina e da necessidade de interfaces voltadas para o usuário final. Os sistemas de recuperação exigem, para isso, modelos de representação do conhecimento que possibilitem contextualizar os significados expressos nos textos armazenados.

É fato que os sistemas de recuperação da informação evoluíram com a utilização de novas tecnologias. No entanto, os resultados são mais visíveis nas interfaces inteligentes e na disponibilização da informação para o usuário final através de redes de comunicação. Em relação ao tratamento do conteúdo, as pesquisas encontram-se ainda em nível experimental. Mesmo assim, são primordiais, uma vez que o tratamento de conteúdo constitui-se no coração do sistema de recuperação da informação. De nada adiantam interfaces inteligentes se elas conduzem à recuperação de documentos irrelevantes, ocasionada por problemas de tratamento de conteúdo.

Há consenso de que quanto mais conhecimento lingüístico/cognitivo for incorporado ao sistema, maior

precisão obter-se-á na recuperação, mas, por sua vez, maior complexidade de implementação e de manutenção. Deve ser considerado, no entanto, que a busca de informação traz implícito o conceito de seletividade e, para isso, o preço pago é esforço, tempo e dinheiro, ou os três juntos, como afirma Meadow [xvi].

Sistemas de recuperação que adotam extração de palavras por meio de métodos estatísticos e aqueles que aplicam análise sintática para extração de sintagmas exigem menor esforço do que os sistemas que incorporam tratamento semântico. Apesar disso, não são capazes de solucionar problemas lingüísticos como a ambigüidade e a sinonímia, tratadas nos sistemas tradicionais que utilizam linguagens documentárias.

Um sistema de recuperação em linguagem natural pode tratar determinados tipos de ambigüidade quando dispõe de informações relativas à valência sintático-semântica das unidades lexicais que compõem um enunciado, como demonstram os resultados dos testes de desambiguação exemplificados neste artigo.
A utilização de Gráficos Conceituais como modelo de representação interna de sistemas de recuperação em linguagem natural pode se constituir em alternativa de solução de ambigüidades que interferem no grau de precisão desses sistemas. A experiência dos sistemas Dr-Link [xvii] e Elen [xviii] demonstram este potencial ao transformar o conteúdo dos documentos e das perguntas dos usuários numa representação em Gráficos Conceituais.

Num modelo desta natureza, a comparação entre o conteúdo dos documentos e da pergunta do usuário efetua-se em nível de conceito - estrutura profunda - e não de forma – estrutura superficial. Como os GCs operam com base em dados sintático-semânticos, possibilitam a interpretação unívoca de formas polissêmicas ou homógrafas.

Esquemas de representações do conhecimento desenvolvidos em outras disciplinas, como Inteligência Artificial, Psicologia e Lingüística, têm despertado interesse crescente na criação de bases de conhecimento que possam ser usadas em recuperação da informação. Cabe aos pesquisadores da área de Ciência da Informação acompanhar os desenvolvimentos dessas áreas e avaliar a possibilidade de aplicação e a adequação de novos métodos e técnicas à recuperação de informação.

NOTAS

[1] A língua oral não é objeto de estudo neste trabalho. Por este motivo, exclui-se aqui a ambigüidade causada por homofonia.

[2] Neologismo sugerido pela Prof. Dra. Enilde Faulstich, orientadora da pesquisa realizada.

[3] Arquivos, no sentido de conjunto de documentos podem ser objeto de compra, mas em contextos muito restritos. Mesmo considerando-se este fator, seria possível desambiguar a frase exemplificada pela análise da relação finalidade, introduzida pela preposição para.

[4] Na teoria dos GCs, as etiquetas de tipo são escritas em letras maiúsculas para diferenciá-las do conceito em si mesmo.

[5] Segundo o número de arcos, as relações conceituais são denominadas por Sowa como monoades (um arco) ; díades (dois arcos) ; tríades (três arcos)...n-ades (n arcos).

[6] Lema: unidade de coleta na sua forma gramatical neutra, i.e., sem conjugação, sem flexão, etc.

[7] Os deverbais, nomes formados a partir de radicais verbais, são denominados por Borba [xi] de nomes abstratos de ação.

[8] O apagamento é uma operação sintática que consiste do cancelamento de um constituinte.[xi].

[9] Ações no sentido de classe conceitual da Ontologia.

REFERÊNCIAS BIBLIOGRÁFICAS

[i] NETWORK Digital Library Thesis and Dissertations.Diponível em:<http://www.ndltd.org>

[ii] OPEN archives initiative. Disponível em:< http://www.openarchives.org >

[iii] CHEN, Hsinchun. Semantic research for digital libraries. D-Lib Magazine, v.5, n. 10 out.1999. Disponível em : <http://www.dlib.org/dlib/october99/chen/10chen.html. > Acesso em: 19 abr. 2001.

[iv] CRANEFIELD, Stephen. Networked knowledge representation and exchange using UML and RDF. **Journal of Digital Information, Southampton**, v. 1, n. 8, fev. 2001. Disponível em: <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Cranefield/>. Acesso em: 12 mar. 2001.

[v] PROCEEDINGS of the workshop on the semantic web: models, architectures and management, Fourth European Conference on Research and Advanced Technology for digital libraries (ECDL 2000). < http://www.ics.forth.gr/proj/isst/SemWeb/proceedins >

[vi] DOERR, Martin. Semantic problems of thesaurus mapping. **Journal of Digital Information, Southampton**, v. 1, n. 8, mar. 2001. Disponível em: <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/>. Acesso em: 12 mar. 2001.

[vii] HUNTER, Jane. MetaNet: a metadata term thesaurus to enable semantic interoperability between metadata domains. **Journal of Digital Information, Southampton**, v. 1, n. 8, fev. 2001. Disponível em: < http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/>. Acesso em: 12 mar. 2001.

[viii] BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The semantic web**. Scientific American, mai. 2001. Disponível em: <http://www.ciam.com/2001/0501issue/0501berners-lee.html >. Acesso em: 19 abr. 2001.

[ix] FUCHS, C. **L'ambiguïté et la paraphrase en linguistique**. In : FUCHS, C., ed. L'ambiguïté et la paraphrase : operations linguistiques, processus cognitifs, traitements automatisés. Caen : Centre de Publications de L'Université de Caen, 1987. p.9 - 20.

[x] FUCHS, C. **Les ambiguïtés du français**. Paris : Orphys, 1996. 183p.

[xi] BORBA, F. S. **Uma gramática de valências para o português**. São Paulo : Ática, 1996. 199p.

[xii] SOWA, J. F. **Conceptual Structures** : information processing in mind and machine.

Massachusetts : Addison-Wesley, 1984. 435 p.

[xiii] WAY, C. E. Conceptual graphs – past, present and future. In : INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES ICCS'94, 2. August 1993, Maryland. **Proceedings...** p. 11-29. (Lectures Notes in Artificial Intelligence, 835).

[xiv] BRÄSCHER, M. **Tratamento automático de ambigüidades na recuperação da informação**. 1999. 286p. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília.

[xv] ZINGLÉ, H. **La modelisation des langues naturelles**: aspects théoriques et pratiques. Travaux du LILLA, numéro spécial, 1999. 151p.

[xvi] MEADOW, C. T. **Text information retrieval systems**. San Diego : Academic Press, 1992. 302p.

[xvii] CHEVALLET, J.-P. **Un modèle logique de recherche d'informations appliqués au formalisme des graphes conceptuels : le prototype ELEN et son expérimentation sur un corpus de composants logiciels**. 1992. Tese (Doutorado) – Université Joseph Fourrier.

[xviii] MYAENG, S. H. ; LI, M. **Linguistic processing of text for a large-scale conceptual information retrieval system**. In: INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES ICCS'94, 2. , August 1994, Maryland. Proceedings...p. 69-83. (Lectures Notes in Artificial Intelligence, 835).

Sobre a autora / About the Author:

Marisa Bräscher
marisa@ibict.br
Doutora em Ciência da Informação pela Universidade de Brasília
Coordenadora Geral de Projetos Especiais do IBICT