UNIVERSIDADE DO EXTREMO SUL CATARINENSE – UNESC

CURSO DE CIÊNCIA DA COMPUTAÇÃO

ADAS PAVEI FONTANA

INTERFACE EM PROCESSAMENTO DE LINGUAGEM NATURAL PARA UROLOGIA

CRICIÚMA, DEZEMBRO DE 2007

ADAS PAVEI FONTANA

INTERFACE EM PROCESSAMENTO DE LINGUAGEM NATURAL PARA UROLOGIA

Trabalho de Conclusão de Curso apresentado para obtenção do Grau de Bacharel em Ciência da Computação pela Universidade do Extremo Sul Catarinense.

Orientadora: Prof. MSc. Merisandra Côrtes de Mattos
Co-Orientador: Prof. MSc. Rozenir Ramos

CRICIÚMA, DEZEMBRO DE 2007

Aos meu pais, Irio e Neuza,

meus irmãos e meus amigos

pelo apoio concedido.

AGRADECIMENTO

Agradeço a Deus por estar presente em todos os momentos da minha vida, confortando nos momentos mais difíceis e alegrando-se nas vitórias.

Agradeço também:

A minha família, por todo o amor, carinho e amizade sempre ajudando na minha formação ética, proporcionando assim uma caminhada sólida e digna.

Ao senhor Candido Batista e sua esposa Pedra Batista que são pessoas que sempre ajudaram e confortaram em momentos difíceis e acima de tudo me orientam a chegar ao caminho certo.

A minha irmã Josiani que está sempre ao meu lado, me ajudando e se dispondo nos momentos da minha ausência.

Ao professor e médico Rozenir Ramos pelo apoio dado na elaboração e construção dos conceitos dispondo sempre seu conhecendo na área de Urologia.

Agradeço a todos os meus amigos que me deram força e sempre estiveram presente na minha caminhada acadêmica.

Agradeço profundamente a grande incentivadora desta pesquisa que não mediu esforços para auxiliar na conclusão da mesma, minha professora e orientadora Merisandra, contribuindo também, além do conhecimento, com sua amizade.

Enfim, a todos que contribuíram para a execução desse trabalho, seja pela ajuda constante ou por uma simples palavra de amizade.

“Tudo o que precisamos aprender,
aprendemos fazendo...”

(Aristóteles)

RESUMO

O uso de tecnologias para o desenvolvimento de interfaces que possam melhorar a interação de usuários com as aplicações têm se destacado no meio computacional, proporcionando, por exemplo, a consulta e o acesso as informações em bases de dados. Dentre essas tecnologias, o uso de processamento de linguagem natural torna-se uma alternativa pelo fato de empregar parâmetros de linguagem utilizados pelos seres humanos na sua comunicação diária. Algumas ferramentas conseguem fazer este tipo de interação para fins educacionais ou em *helpdesk* de organizações, entre outros. Esta pesquisa fundamentou-se no processamento de linguagem natural, por meio do reconhecimento de sentenças, criando-se uma interface de consulta, desenvolvida em linguagem Java na IDE Netbeans 5.5, denominada de Hades, que pode ser utilizada por estudantes de Medicina e profissionais da área de urologia. Na realização dos testes do Hades verificou-se que aplicando processamento de linguagem natural em interfaces podem-se disponibilizar outras formas de interação homem-computador, por vezes mais simples, obtendo-se assim resultados satisfatórios.

Palavras-Chaves: Inteligência Artificial, Processamento de Linguagem Natural, Interface em Linguagem Natural, Informática em Saúde, Urologia.

ABSTRACT

The use of technologies for the development of interfaces that you can improve the users' interaction with the applications has if outstanding in the half computational, providing, for instance, the consultation and the access the information in bases of data. Among those technologies, the use of processing of natural language becomes an alternative for the fact of using language parameters used by the human in your daily communication. Some tools get to do this interaction type for educational ends or in helpdesk of organizations, among others. This research was based in the processing of natural language, through the recognition of sentences, growing up a consultation interface in language Java in the you IDE Netbeans 5.5, denominated of Hades, which can be used by students of Medicine and professionals of the urology area. In the accomplishment of the tests of Hades it was verified that applying processing of natural language in interfaces availability other forms of interaction man-computer is been able to, for simpler times, obtaining like this satisfactory results.

Keu-Words: Intelligence Artificial, Natural Language Processing, Interface in Natural Language, Information in the Health, Urology.

LISTA DE FIGURAS

LISTA DE SIGLAS

ART	Artigo
CID	Código Internacional de Doenças
CID-10	Código Internacional de Doenças versão 10
GLC	Gramática Livre de Contexto
HPB	Hiperplasia Prostática Benigna
IA	Inteligência Artificial
LN	Linguagem Natural
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
PEP	Prontuário Eletrônico do Paciente
RES	Registro Eletrônico em Saúde
SNOMED	Systematized Nomenclature of Medicine
SN	Sintagma Nominal
SUBS	Substantivo
SUS	Sistema Único de Saúde
SV	Sintagma Verbal
UML	Unified Modeling Language

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é uma área de estudo da Inteligência Artificial (IA) que tem por objetivo dotar as interfaces de computadores da capacidade de comunicar-se com seu usuário no idioma deste. Portanto, o processamento de linguagem natural expressa o que é construído por meio da linguagem, sendo esta a principal forma de interação entre as pessoas. Podendo, assim, ser aplicado a vários tipos de componentes de softwares, como por exemplo, nas consultas em banco de dados, busca de informação, entre outros (HÜBNER, 1992).

A técnica de PLN tem sido aplicada em diferentes áreas do conhecimento, dentre estas a da saúde. Esta utilização deve-se a necessidade de uma interface que possibilite uma maior interação entre o sistema computacional e o médico.

Entre outras aplicações, a informática em saúde pode ser utilizada para o registro informatizado dos dados do paciente, sendo empregada para dar suporte aos usuários (profissionais de saúde) por meio da disponibilidade de informações completas e confiáveis, lembretes e alertas, sistemas de apoio à decisão, atalhos para bases de conhecimentos virtuais, entre outros (IOM, 1997). Assim, podem ser geradas bases de dados a partir de prontuários eletrônicos, como por exemplo, na área de urologia.

A urologia é uma especialidade da Medicina que previne e trata doenças referentes ao sistema urinário de homens, mulheres e crianças, bem como do sistema reprodutor masculino.

Nesta pesquisa utilizou-se uma base de dados da área de urologia a fim de se desenvolver o protótipo de um sistema com interface em linguagem natural, denominado de Hades, onde o médico pode consultar informações do paciente

usufruindo da própria linguagem utilizada em um diálogo. A forma de consulta sobre a condição do paciente ocorre por meio de questionamentos, como por exemplo:

Médico: Qual paciente possui hiperplasia prostática benigna?

Hades: 1058.

Este tipo de consulta busca melhorar a utilização dos sistemas, já que para grande parte das pessoas a interação com os recursos disponíveis, na solução de seus problemas, causa receio pela diversidade de interfaces gráficas (HÜBNER, 1992).

Assim, a fim de projetar a interface do Hades foram adquiridos, processados e demonstrados os conhecimentos referentes a base de dados da área de urologia, pois a interface em processamento de linguagem natural se mostra muito parecida com o processo de aquisição do conhecimento humano.

A utilização de PLN no Hades consistiu na compreensão da linguagem basicamente por meio de quatro métodos: análise léxica, sintática, semântica e pragmática. Estas análises permitem dividir, reconhecer e responder a uma questão, do domínio de aplicação, gerada pelo usuário em um idioma específico.

1.1 OBJETIVO GERAL

Desenvolver uma interface por meio de processamento de linguagem natural para consulta a uma base de dados de Urologia.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste projeto são:

a) compreender o processamento de linguagem natural;

b) utilizar a técnica de processamento de linguagem natural em uma interface de consulta;
c) aplicar os métodos de processamento de linguagem natural: análise léxica, sintática, semântica e pragmática;
d) empregar uma base de dados da área de urologia;
e) proporcionar uma ferramenta inteligente para consultas em uma base de dados da área de urologia.

1.3 JUSTIFICATIVA

A interação entre o homem e a máquina pode ocorrer por meio de interfaces gráficas com menus, ícones e também pelo uso de linguagem natural. Dentre estas, os menus que criam uma hierarquia de comandos são os mais utilizados, porém podem se tornar cansativos quando há muitos níveis. Enquanto, a linguagem natural torna-se simples ao usuário, pois proporciona uma comunicação semelhante a forma habitualmente empregada pelos seres humanos, possibilitando o estabelecimento de diálogos quando necessário (HÜBNER, 1992; RODRIGUES,1998).

O processamento de linguagem natural deve possuir características dos seres humanos como: aquisição, retenção e exposição do conhecimento, sendo estas tarefas de estudo da IA, proporcionando aos sistemas a utilização em diferentes áreas de atuação. Assim, o processamento de linguagem natural está intimamente ligado às capacidades mentais humanas (HÜBNER, 1992).

Nesta definição, o processamento de linguagem natural se agrega a vários tipos de sistemas, como consultas em banco de dados e também operacionalizando

sistemas médicos por meio da utilização de bases de dados de especialidades como urologia.

No que se refere à urologia de acordo com o Instituto Nacional do Câncer (INCA) (2006) a principal doença dessa especialidade é o câncer de próstata que atinge em uma amostra de 100.000 0,51%, sendo que 75% dos casos ocorrem após os 65 anos de idade.

A base de dados da área de urologia, utilizada nesta pesquisa, contém informações de pacientes, principalmente do sexo masculino e de doenças em seu sistema reprodutor. Esta base de dados foi disponibilizada pelo médico urologista Rozenir Ramos, professor do curso de Medicina da Universidade do Extremo Sul Catarinense.

Esta pesquisa aplicou o PLN em uma interface de consulta a essa base de dados, podendo proporcionar uma busca por contexto. Conforme Wives (1996) a utilização de PLN na área da saúde pode determinar porque alguns pacientes apresentam melhora rapidamente, ou até mesmo diagnosticar mais facilmente determinada doença analisando-se seus sintomas.

Além disso, o médico não ficará restrito aos métodos usuais de uma linguagem técnica programada para o sistema, passando a escrever suas consultas em linguagem natural (SCHIPIURA, 2004).

O processamento de linguagem natural visa a utilização de análises que fazem a compreensão da frase e produzem respostas para a consulta. A realização destas análises será de forma léxica, sintática, semântica e pragmática, decompondo e analisando o sentido da frase no contexto da área médica.

Dessa forma, esta pesquisa visa contribuir na criação de uma interface de consulta a uma base de dados da área de urologia, podendo auxiliar na operacionalidade

desta e proporcionar uma interface mais intuitiva, baseando-se no uso da linguagem. Modelos computacionais como estes que aliam recursos de inteligência artificial, podem facilitar os processos de busca e compreensão do conhecimento, disponibilizando de forma simples e em linguagem compreensível pelo usuário respostas para alguns argumentos médicos.

1.4 ESTRUTURA DO TRABALHO

Esta pesquisa é composta por sete capítulos. No Capítulo 1 encontra-se a contextualização ao tema proposto, bem como os objetivos e a justificativa para realização deste trabalho.

Os conceitos fundamentais de Linguagem são abordados no Capítulo 2, realizando-se uma introdução ao processamento de línguas naturais.

No Capítulo 3 é apresentado os conceito e as análises envolvidas no Processamento de Linguagem Natural, que é essencial para o entendimento da pesquisa realizada, bem como os tipos de aplicações com ênfase nas características de interface em linguagem natural

O Capítulo 4 apresenta informações na área da saúde, pontuando conceitos e tipos. Já no Capítulo 5 são abordados alguns exemplos de aplicações da área da saúde que usam processamento de linguagem natural.

O Hades, a base utilizada bem como todos os procedimentos de análises e resultados obtidos na interface são apresentados no Capítulo 6.

Por fim, tem-se a conclusão, onde também são descritas sugestões para os trabalhos futuros.

2 A LINGUAGEM E A INTELIGÊNCIA ARTIFICIAL

A linguagem exerce um fascínio sobre a humanidade, não só pela capacidade de criação e transformação, mas também por possibilitar o compartilhamento de experiências. Sendo assim, linguagem é qualquer conjunto de sinais que um indivíduo utiliza para se comunicar (FIORIN, 2003).

Toda e qualquer linguagem traz com o ser humano partes próprias do meio onde ele nasceu e cresceu, podendo-se alterar em função da sua convivência com outras pessoas em diferentes ambientes.

Ao utilizar a linguagem, o indivíduo vale-se do idioma em que aprendeu a falar e a escrever para se comunicar, pois, possui um número infinito de frases e se mostra um componente importante neste processo. O idioma é uma característica particular, sendo diferente em vários países. Cada comunidade possui seu idioma, mas todas têm como parte integrante do processo de comunicação o alfabeto (LANGACKER, 1980).

O alfabeto é um conjunto finito de símbolos, que consiste em uma entidade abstrata básica, tendo-se como exemplo as letras (MENEZES, 2005). O alfabeto compõe o idioma que é denominado de linguagem natural.

2.1 LINGUAGEM NATURAL

A linguagem natural é utilizada para a comunicação entre os indivíduos, sendo de fundamental importância no entendimento de mensagens que são enviadas diariamente e encontra-se nos diferentes lugares em que o homem vive proporcionando a reflexão e expressão dos pensamentos (FIORIN, 2004).

O ser humano se comunica utilizando a linguagem natural, mas para isso ele deve estar ciente que ela possui uma estrutura, sendo composta por regras que padronizam a comunicação, tornando-a compreensível por todos.

2.1.1 Estrutura da Linguagem Natural

Consideram-se as seguintes palavras: casa, casinha e casebre, todas estas são formadas por mais de um elemento, sendo comum a todas elas "*cas*", denominando de radical da palavra. Os outros elementos compõem e modificam o radical e ocasionam à derivação da mesma, por meio de afixos, flexão de número, gênero e grau.

Na compreensão de uma frase em linguagem natural podem ocorrer vários tipos de situações como:

a) **ambigüidade**: consiste em deixar a frase com mais de um sentido (TERRA, 1997). Exemplo: "A *excelente* Dona Inácia era mestra na arte de judiar crianças" (LOBATO, 1982, p. 50). A palavra *excelente* expressa ambigüidade, pois seu significado é alterado no contexto da frase;

b) **anáfora**: a repetição da mesma palavra no início de dois ou mais versos chama-se anáfora. Esse recurso, na linguagem poética, é perfeitamente possível, como por exemplo, no texto a seguir de Fernando Pessoa, onde tem-se a repetição da palavra *serei*. Na linguagem formal (dissertação) a repetição abusiva pode não levar ao mesmo resultado (SACCONI, 1992);

> *Serei* sempre o que não nasceu para isso;
> *Serei* sempre o que tinha qualidades;
> *Serei* sempre o que esperou que lhe abrissem a porta
> ao pé de uma parede sem porta. (PESSOA, 1990)

c) **ironia**: utilização de palavras ou frases que divergem do sentido literal, constituindo-se em sátira[1] ou conotação[2] (TERRA, 1997). Exemplo: “o pobre homem perdeu, as estribeiras; ficou cego de ciúme.” Na expressão *pobre homem* ocorre um caso de duplicidade do significado da frase, pois se pode entender que é um homem pobre ou um homem cansado.

Conforme Terra (1997) a linguagem natural está voltada, essencialmente, a três aspectos da comunicação: fonologia, estrutura e significado.

A fonologia diz respeito aos sons emitidos que integram as frases, como a linguagem natural é uma área abrangente em termos de conhecimento, serão abordados com mais relevância os aspectos estruturais e de significado das palavras, pois compreendem o foco desta pesquisa.

Estruturalmente a frase se divide em morfologia e sintaxe. A parte morfológica reconhece a palavra na sua forma mais primitiva. Considerando-se, por exemplo, a palavra apreensão tem-se: apreen + são

radical sufixo

Na sintaxe da frase as palavras se relacionam logicamente entre si (TERRA, 1997). Na Figura 1 há uma estrutura sintática em forma de árvore.

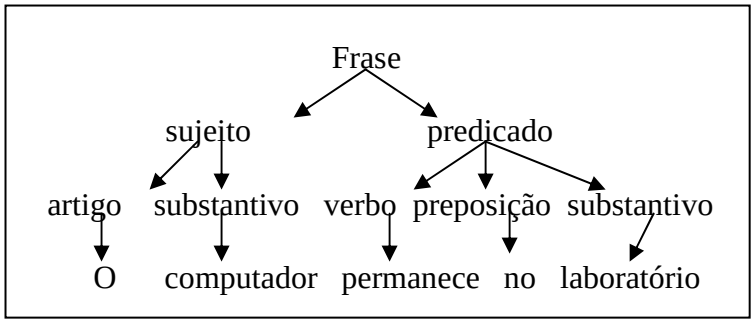

Figura 1. Composição da Frase

[1] É uma forma literária de ridicularizar um determinado tema ou indivíduo (TERRA, 1997)
[2] Nos textos literários a linguagem nem sempre apresenta um único sentido, quando este for diferente daquele que lhe é comum chama-se conotação (TERRA, 1997)

Após a verificação das palavras dentro da gramática procura-se identificar o contexto da estrutura sintática, em termos dos significados das palavras que a compõem, denominando-se de semântica (TERRA, 1997). Na Figura 1, por exemplo, entende-se que o objeto persiste, ou seja, está em algum lugar.

E, finalmente a análise pragmática verifica se o significado de uma ou qualquer estrutura sintática está de acordo com o contexto da frase. Exemplo: persiste/está = permanece.

A análise de linguagem natural a partir da sua estrutura, fundamenta aspectos que podem ser utilizados dentro da área computacional por meio do PLN, compreendendo a análise léxica, sintática, semântica e pragmática. Estes aspectos da linguagem natural são abordados pela área de Inteligência Artificial, que busca métodos computacionais para desenvolver sistemas inteligentes.

2.2 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é uma área de estudo da computação que permite analisar e criar sistemas com comportamento inteligente realizando tarefas complexas com um nível de competência que se equipara a um especialista humano, buscando para isto, tudo que envolva aprendizado (LUGER, 2004).

Os problemas que a IA resolve são relacionados ao contexto de utilização e trazem respostas que se adequam ao sistema. Assim, é preciso buscar o conhecimento sobre o problema/causa que se mostra fundamental para a resolução e construção de sistemas (COELHO, 1995).

A IA abrange, como a maioria das ciências, uma série de técnicas que compõem e geram aplicações com diferentes abordagens:

a) **representação de conhecimento:** refere-se a representação na base de conhecimento dos aspectos críticos da atividade inteligente para desenvolvimento de um sistema computacional. Esta base realiza o mapeamento entre os objetos e as relações de um domínio de aplicação com os de um programa (LUGER, 2004);

b) **raciocínio automatizado**: é a simulação computacional do processo de pensar, realizado pelo indivíduo. A importância do raciocínio automático deve-se a liberação do homem de processos repetitivos, rotineiros e cansativos, para que ele se dedique a tarefas que necessitem de um certo nível de criatividade, atualmente, um tanto quanto distantes da máquina (RUSSEL; NORVIG; LUGER, 2004);

c) **aprendizado de máquina**: é um dos componentes mais importantes no comportamento inteligente. O computador dota-se de inteligência e precisa conhecer a linguagem natural, pensamento e aprendizado. O aprendizado torna a máquina apta a novos conhecimentos (LUGER, 2004);

d) **visão computacional:** é um campo da IA que está em crescimento, mas ainda no início de seus estudos, constituindo-se do conjunto de métodos e técnicas por meio dos quais sistemas computacionais podem ser capazes de interpretar imagens (WANGENHEIM, 2003);

e) **robótica**: é uma área que envolve várias outras como: engenharia, mecânica, eletrônica, física, ente outras, tratando da composição de sistemas feitos por meio de máquinas e mecânica automatizada, sendo controlado por circuito integrado. Assim, a robótica produziu muitas

idéias que dão suporte na solução e problemas orientados a agentes (RUSSEL; NORVIG, 2004);

f) **processamento de linguagem natural**: compreensão de uma seqüência de símbolos e geração de uma outra (linguagem natural – textos). A tarefa de processar uma linguagem natural permite que os seres humanos comuniquem-se com os computadores de uma forma habitual, utilizando-se para isso da linguagem com a qual estão mais acostumados (COSTA; SIMÕES, 2004).

3 PROCESSAMENTO DE LINGUAGEM NATURAL

O processamento de linguagem natural relaciona-se a área computacional, bem como a outras ciências, como por exemplo, a lingüística e psicologia. A lingüística é uma ciência que visa caracterizar e explicar a multiplicidade de formas das expressões, por meio de conversas, textos escritos, bem como de outros meios de comunicação. Compreende também a aquisição, produção e compreensão da linguagem pelos seres humanos e das relações existentes entre esta e o mundo (SCHÜTZE; MANINNG, 1999).

Na psicologia há um espaço especial reservado à pesquisa orientada em sentido lingüístico, ou seja, referente ao comportamento verbal. Assim, o estudo da linguagem voltada à psicologia (psicolingüística) refere-se a codificação e decodificação do processo de comunicação, onde ligam-se os estados das mensagens e dos comunicantes (TITONI, 1983).

No que tange a área computacional, esta técnica da IA consiste em um conjunto de métodos para realizar análises textuais e gerar frases em qualquer idioma compreensível pelo ser humano (KRULEE, 1991).

O PLN tem como objetivo habilitar os computadores com a capacidade de analisar e interpretar a linguagem humana, processando-a e fornecendo uma resposta, como no caso dos sistemas de consultas, tradução automática ou de criação de sumários de textos (TANIWAKI, 2001).

A análise e o entendimento de um texto, por meio de PLN, significam reconhecer o seu contexto, realizar as análises sintática, semântica e léxica, criar resumos, extrair informação, interpretar os diferentes sentidos das palavras e aprender novos conceitos com os textos processados (ALEXANDRINI, 2005).

Todas essas tarefas do PLN são essenciais para o computador entender comandos escritos em linguagem natural, possuindo etapas parecidas com a compilação de linguagem de máquina (Java, C, Pascal), porém voltadas a realização da compreensão do idioma (TANIWAKI, 2001).

Segundo Allen (1994), as aplicações em processamento de linguagem natural podem ser baseadas em textos ou diálogos. As baseadas em textos envolvem processamento das tarefas para extração de informações como, por exemplo, de mensagens ou textos. As aplicações voltadas para o diálogo buscam a comunicação homem-computador usando a fala ou a escrita, envolvendo sistemas de perguntas e respostas. Esta abordagem é bastante utilizada em *helpdesks*[3] de grandes organizações, comandos por meio de voz, entre outras aplicações.

Aplicações como estas que envolvem PLN, estão se desenvolvendo ao longo dos anos e buscam aprimorar a interação do homem com o computador. Os primeiros trabalhos desenvolvidos datam da década de 50 e se dedicavam ao tratamento informatizado da linguagem natural (ALEXANDRINI, 2005).

Um dos primeiros programas de computador que utilizou linguagem natural, denominado Eliza[4], foi desenvolvido no Instituto de Tecnologia de Massachusetts por Joseph Weizenbaum em 1966 (HUBNER, 1992). O sistema conduz uma conversa com o usuário no estilo de um psicólogo Rogeriano[5]. Na Figura 2 e 3 ilustra-se uma conversa, por meio de afirmações que interrogam o usuário, e assim sucessivamente, como se fosse um psicólogo.

[3] Designa o serviço de apoio à usuários para suporte e resolução de problemas técnicos em informática, telefonia e tecnologias de informação (MORAES, 2002)
[4] Disponível em http://www-ai.ijs.si/eliza/eliza.html
[5] Realiza perguntas que levam o paciente a descobrir seus problemas por meio das suas próprias palavras (HUBNER, 1992).

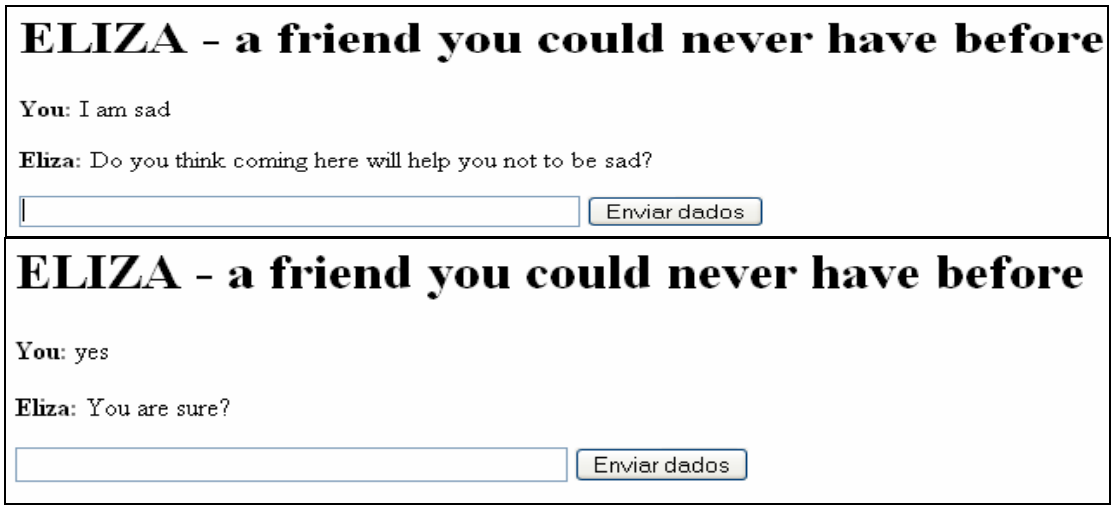

Figura 2. Diálogo com ELIZA.
Fonte: HUBNER (1992)

No Brasil, a primeira experiência no campo de PLN foi um programa *web* chamado Cybelle que tem estrutura semelhante ao ELIZA (CANTARELLI, 1998), que simula uma conversa humana em idioma português sobre diferentes assuntos. Na Figura 3 há um exemplo de diálogo entre o usuário e o *chatterbot*[6] Cybelle[7]. O Cybelle foi desenvolvido pelo estudante de doutorado Alex Primo da Universidade Federal do Rio Grande do Sul e o programador Luciano Coelho que constataram em meados de 1999 que não havia um *chatterbot* em português.

[6] São programas de computador que tentam simular conversações com os usuários, com o objetivo de temporariamente levar um indivíduo a pensar que está conversando com outra pessoa (CANTARELLI, 1998).
[7] Disponível em http://cybelle.cjb.net/.

Figura 3. Diálogo com o *chatterboot* Cybelle
Fonte: ALEXANDRINI (2005)

Outros *chatterbots* estão disponíveis na internet, principalmente direcionados para o atendimento via *web* ao cliente (sugestões, reclamações e explicações de produtos), como também para a busca de conteúdo. Um exemplo deste tipo de pesquisa (Figura 4) encontra-se no *site* http://www.msdewey.com/ que disponibiliza uma bibliotecária que interage com o usuário enquanto faz suas pesquisas sobre diversos assuntos.

Figura 4. Chatterbot MsDewey
Fonte: MICROSOFT (2007)

Além dos *chatterbots*, têm-se outros tipos de aplicações voltadas a recuperação e extração de informação, bem como a tradução automática. A recuperação de informação consiste em recuperar documentos ou informações contidas nele. Já a extração busca entradas em banco de dados para procurar em um texto a ocorrência de um objeto, como por exemplo, prever tempestades por meio de relatórios meteorológicos. Na tradução automática, aplicação mais tradicional e utilizada de PLN, ocorre a conversão de um texto em uma linguagem para outra. A tradução é um processo complexo por exigir uma maior compreensão do texto e da situação que está sendo comunicada (RUSSEL; NORVIG, 2004).

Os sistemas de linguagem natural utilizam um conhecimento considerável sobre a estrutura da linguagem, incluindo o que são as palavras, como elas combinam para formar frases, o que significam como contribuem para o entendimento da frase, e assim por diante. Além disso, deve-se também considerar que o homem, como um ser inteligente, possui um conhecimento geral e uma habilidade de raciocínio, por exemplo, em uma conversação a pessoa não está somente envolvida nos aspectos estruturais da linguagem usada, mas também no conhecimento geral da conversa e seus aspectos particulares (ALLEN, 1994).

3.1 ESTRUTURA DO PROCESSAMENTO DE LINGUAGEM NATURAL

Levando-se em consideração o processo de estruturação da frase realizado em relação à palavra, sentido e contexto, o processamento de linguagem natural divide-se em algumas etapas para a compreensão correta da oração dentro do conceito computacional. A primeira refere-se à análise léxica que identifica palavras ou expressões isoladas em uma sentença, sendo este processo auxiliado por delimitadores

(pontuação e espaços em branco). As palavras identificadas são classificadas de acordo com a sua utilização ou, no caso da linguagem natural, categoria gramatical.

Na segunda etapa, por meio da gramática da linguagem e das informações do analisador morfológico, a parte sintática procura construir árvores de derivação[8] para cada sentença, mostrando como as palavras estão relacionadas entre si. A análise semântica é a terceira etapa que estuda a atribuição do significado para as frases e palavras, neste caso, na língua portuguesa. A quinta e última etapa compreende a análise pragmática que não verifica o significado das partes, mas sim a interpretação no contexto da frase.

3.1.1 Análise Léxica

A análise léxica é a primeira fase do PLN, sendo responsável pela identificação das palavras e sua classificação conforme o tipo (artigo, substantivo, verbo, entre outros). Esta verificação é usada como entrada em um sistema de linguagem natural (PRICE; TOSCANI, 2000).

Conforme Hubner (1992) o analisador léxico recebe e lê o texto de entrada, realizando algumas tarefas em nível de interface com o usuário. A principal delas consiste na remoção dos espaços em branco da entrada, como por exemplo, na frase: “Qual a incidência de câncer de próstata nos pacientes?”. No momento em que o analisador léxico realiza a retirada dos espaços em branco e caracteres sem utilidade, decompõe a frase por palavra, ficando da seguinte forma: [qual][a][incidência][de][câncer][de][próstata][nos][pacientes].

[8] É a representação gráfica da derivação de uma sentença particular, de acordo com a gramática especificada (KRULEE, 1991).

Depois que a frase é analisada, as palavras são verificadas dentro de um dicionário de dados para dar-se a classificação gramatical correspondente.

Os dicionários de dados fazem parte da maioria dos programas de linguagem natural sendo concebido para utilização com codificação lingüística compreensível ao sistema na realização da análise léxica. As palavras do dicionário estão ligadas as regras gramaticais do idioma, portanto se o sistema não encontrar a palavra processada gera-se um erro, como por exemplo, a palavra não existe no dicionário ou está escrita de forma incorreta. Caso a palavra esteja correta o processamento continua identificando a frase e a dividindo em *tokens*[9] (WIVES, 1996).

Nesta fase o analisador lê uma cadeia de caracteres coletando os *tokens* que consistem nas palavras em linguagem natural utilizadas como entrada no programa. Esses *tokens* são unidades que compõem a frase, constituindo-se na parte mais básica do PLN (MENEZES, 2005).

Considerando-se a língua portuguesa, por exemplo, as palavras podem ser classificadas como (TERRA, 1997):

a) **substantivo:** nomeia os seres em geral como animais, lugares, pessoas, entre outros. Exemplo: carro, Brasil, Deus, saudade;
b) **adjetivo:** caracteriza o substantivo, indicando qualidade, estado e modo de ser. Exemplo: limpo, vazio, bom;
c) **artigo:** precede o substantivo, indicando gênero e número; ao mesmo tempo determina-o de modo vago ou preciso. Exemplo: o, a, um, uma;
d) **numeral:** indica quantidade ou ordem de sucessão. Exemplo: um, dois, terceiro, dois terços;

[9] Armazena cada palavra, bem como a sua classificação na gramática de acordo com o idioma adotado. (TERRA, 1997)

e) **pronome:** representa ou acompanha o substantivo, indicando-o como pessoa do discurso. Exemplo: ele, sua, minha, meu;

f) **verbo:** exprime um fato como ação, estado ou fenômeno situando-no tempo. Exemplo: estudar, conhecer, amar;

g) **advérbio:** modifica o verbo, o adjetivo ou ainda outro advérbio, demonstrando determinada circunstância. Exemplo: bastante, certamente, calmamente;

h) **preposição:** relaciona dois termos desta, subordinando ao outro. Exemplo: de, sobre, com, após;

i) **conjunção:** relaciona orações ou termos da oração que exercem a mesma função. Exemplo: Tomei café. Fui ao trabalho. Tomei café **e** fui ao trabalho;

j) **interjeição:** denota emoções súbitas. Exemplo: Nossa! Puxa! Socorro!

Essas classes gramaticais são partes integrantes de uma frase, que constroem um sentido completo ao discurso. Com isso, as palavras podem as vezes mudar sua classe gramatical, devido ao contexto em que são empregadas. Por exemplo:

Há coisas **úteis** e coisas inúteis.
Adjetivo

Devemos unir o **útil** ao agradável.
Substantivo

Essa alternância pode ser denominada de ambigüidade e constitui-se em um dos grandes desafios do PLN, pois a mesma palavra pode mudar sua classe gramatical de acordo com o contexto da frase.

Assim, o reconhecimento da palavra passa por um processo de leitura seguindo um estágio de *tokenização* destas e referenciando seus atributos. Dessa forma,

o analisador recebe uma entrada x (caractere) e armazena em *buffer* até o momento que encontrar um espaço em branco, formando assim uma cadeia de caracteres (palavra).

O próximo passo é verificar se essa palavra está inserida no dicionário da língua portuguesa e a qual classe pertence. Finalizando, o analisador léxico pode identificar possíveis erros referentes a escrita da palavra ou inexistência dela no dicionário de dados. A seguir, realiza-se a próxima fase do PLN chamada de análise sintática (Figura 5).

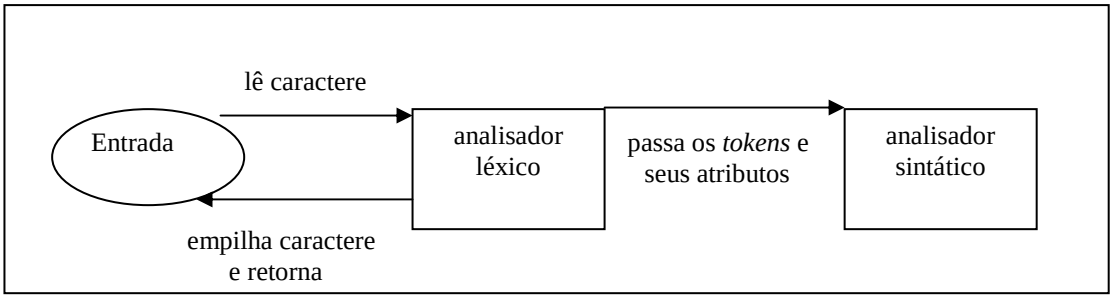

Figura 5. Exemplo de entrada de dados.

3.1.2 Análise Sintática

Após o reconhecimento das palavras pelo analisador léxico, a análise sintática determina a estrutura da sentença, por meio da gramática que define a organização lingüística. Dessa forma, as palavras devem se relacionar gramaticalmente na frase (MENEZES, 2005).

A estrutura de uma sentença está presente no conjunto de palavras que originam:

a) **oração:** enunciado constituído de sujeito e predicado ou somente deste. A oração se organiza a partir do verbo. Por exemplo: “Parecia um deus sentado naquela poltrona”;

b) **frase:** enunciado lingüístico que possui sentido completo e não necessariamente se organiza em torno de um verbo. Exemplo: “Silêncio!”

c) **período:** qualquer frase que pode ser constituída por uma ou mais orações. Exemplo: “Amor é vida; é ter constantemente”

“Alma, sentidos, coração – abertos”

Na composição das frases, orações e períodos são empregadas regras gramaticais que proporcionam a correta utilização das palavras. Portanto, entende-se que esta compreende a construção de uma derivação ou árvore sintática para uma cadeia de entrada a partir da definição formal de uma gramática (LUGER, 2004).

3.1.2.1 Gramática

Gramática é uma estrutura de análise que verifica por meio de um conjunto de regras uma determinada composição da frase. Basicamente, gramática é um conjunto finito de regras as quais quando aplicadas sucessivamente, geram palavras e o conjunto destas define a linguagem (MENEZES, 2005).

Uma gramática qualquer é composta por uma quádrupla ordenada: G=(N,T,P,S). Onde:

a) N é o conjunto de símbolos não-terminais (variáveis);

b) T é o conjunto de símbolos terminais (constantes);

c) P é o conjunto de regras (regras sintáticas);

d) S é o símbolo inicial da gramática.

Como exemplo de quádrupla, tem-se a Figura 6, onde o conjunto N (não terminais) é composto por *frase, sujeito, predicado, artigo, substantivo* e *verbo* (estes

símbolos compõem a classe gramatical de cada palavra), o conjunto T (terminais) são *o*, *a*, *carro, moto, bater* e *correr* e por fim o símbolo inicial desta gramática é composto pelo não terminal *frase*. Sendo assim, a gramática possui seis regras sintáticas estabelecendo por meio do símbolo inicial o princípio da gramática.

Frase-> sujeito|predicado

Sujeito-> artigo|substantivo

Predicado-> verbo|artigo|substantivo

Artigo-> o|a

Substantivo-> carro|moto

Verbo-> bater|correr

Figura 6. Exemplo de Gramática.

Levando-se em consideração este tipo de estrutura, uma gramática utilizada em PLN pode ser divida em: Sensível ao Contexto e Livre de Contexto.

3.1.2.1.1 Gramática Livre de Contexto

A Gramática Livre de Contexto (GLC) possui regras de produção P e apresenta a seguinte representação: $X \rightarrow \alpha$, onde X pertence ao conjunto dos N e α ao conjunto (N U T). Levando-se em consideração esta definição, permite que as regras possuam apenas um único símbolo não-terminal no seu lado esquerdo (MENEZES, 2005). Consequentemente, conforme Luger (2004), a regra pode ser aplicada a qualquer ocorrência daquele símbolo de forma independente do seu contexto.

Assim, entende-se que o termo livre de contexto refere-se ao fato das variáveis (não-terminais) X poderem ser substituídas por α em todo e qualquer contexto (PRICE; TOSCANI, 2000).

Na Figura 7 tem-se um exemplo de uma gramática livre de contexto. A partir desta, pode-se gerar a seguinte oração: *O carro bateu na moto.*

Frase-> sujeito|predicado

Sujeito-> artigo|substantivo

Predicado-> verbo|preposição|substantivo

Artigo-> o

Preposição -> na

Substantivo-> carro|moto

Verbo-> bateu|correu

Figura 7. Gramática Livre de Contexto.

As GLC são utilizadas em grande parte das implementações de PLN, devido ao fato de apresentar um desempenho melhor em relação as outras, sendo a sua implementação mais simples. Além disso, possui alguns bons requisitos para o PLN, como a não utilização de terminais no seguinte estilo: aX → aaXb, o que diminui a complexidade de análises das produções na gramática (HÜBNER, 1992).

3.1.2.1.2 Gramática Sensível ao Contexto

As gramáticas sensíveis ao contexto permitem mais de um símbolo no lado esquerdo de uma regra e possibilitam definir o contexto em que essa regra é aplicada. Assim, assegura-se a satisfação de exceções, como por exemplo, concordância em

número e outras verificações semânticas. As regras sensíveis ao contexto possuem a restrição de que o lado direito seja tão longo quanto o esquerdo (LUGER, 2004).

Esta gramática caracteriza-se pelas produções P possuírem o formato (KRULEE, 1991):

$$\gamma_1 \, X \, \gamma_2 \rightarrow \gamma_1 \, \alpha \, \gamma_2$$

Onde:

a) X pertence aos N (símbolos não-terminais);

b) γ_1, γ_2 e α estão no conjunto (N U T). Sendo que, somente X pode ser restrito a α em um contexto $\gamma_1 \gamma_2$.

Na Figura 8 tem-se o exemplo de uma gramática que gera uma linguagem qualquer, podendo-se originar as sentenças presentes na Figura 9. Nestas gramáticas sensíveis ao contexto deve-se considerar que as regras só podem ser utilizadas em contexto onde o lado-esquerdo, incluindo N e T, coincide com o conjunto que se deseja provar.

S → abc | aX1bc

X1c → X2bcc

aX2 → aaX1 | aa

bX2 → X2b

X1b → bX1

Figura 8. Gramática Sensível ao Contexto.
Fonte: KRULEE, G. (1991)

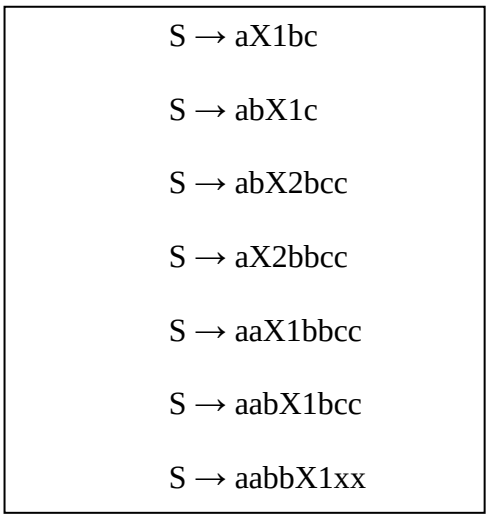

Figura 9. Sentenças
Fonte: KRULEE, G. (1991)

As gramáticas sensíveis ao contexto podem definir estruturas de linguagem que não são identificadas quando se utiliza as livres de contexto, porém possui uma série de desvantagens para a elaboração de analisadores sintáticos, dentre as quais LUGER (2004) destaca:

a) aumentam o número de regras e símbolos não terminais da gramática a fim de incluir concordâncias de gênero, número e outras necessárias, por exemplo, pela língua portuguesa;

b) dificultam o entendimento da estrutura frasal da linguagem.

Além da gramática, usa-se em PLN uma derivação de regras por meio de árvores de derivação. Assim, a análise sintática gramatical pode ser descrita como um procedimento que busca diferentes maneiras de combinar regras gramaticais gerando uma árvore que representa a estrutura da frase.

3.1.2.2 Árvore de Derivação

Nas aplicações de PLN é conveniente representar a derivação de palavras na forma de árvore, partindo do símbolo inicial como a raiz e finalizando em símbolos terminais chamados de folhas (MENEZES, 2005).

A derivação em uma GLC prevê a definição com um S_0 que representa o nó inicial da árvore e a partir desse são produzidas sentenças, criando-se ramos que são símbolos não-terminais (X,Y). Os ramos X,Y por sua vez podem ter ramificações até chegaram aos nós folhas, ou seja, os símbolos terminais da gramática (Figura 10) (KRULEE, 1991).

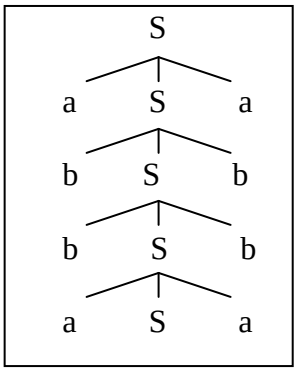

Figura 10. Árvore de Derivação.
Fonte: KRULEE, G. (1991)

Na Figura 10 tem-se a árvore de derivação em uma gramática livre de contexto, porém pode-se utilizá-la em outros tipos de gramáticas (KRULEE, 1991).

Uma combinação gramatical pode gerar dois tipos de derivação, conforme o algoritmo busca uma sentença chave, podendo ocorrer de um ponto da gramática e descer até os nós finais (top-down), ou parte desses e vai até o ponto de início (botton-up).

A derivação *top-down* começa com o símbolo inicial da gramática (S_0) e tenta reescrevê-lo introduzindo uma seqüência de símbolos terminais que combinam com as classes de palavras na sentença de entrada. Isso significa que a partir de S_0

aplica-se a regra $S_0 \rightarrow$ SN SV, sendo a lista de símbolos composta por Sintagma Nominal[10] (SN) e Sintagma Verbal[11] (SV). A partir do SN deriva-se Artigo (ART) e Substantivo (SUBS), tendo-se então a regra $S_0 \rightarrow$ ART SUBS SV e assim sucessivamente até os nós terminais da gramática (Figura 11)(ALLEN, 1994).

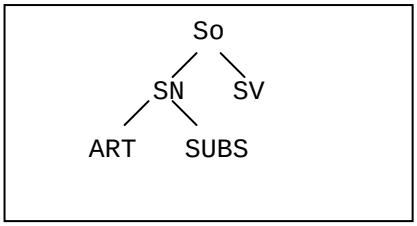

Figura 11. Derivação *Top-Down*

A derivação *bottom-up* começa a partir dos nós filhos em direção ao símbolo inicial da gramática. Dessa forma, usa-se a regra para obter a seqüência ART e SUBS que encontrou e identifica como um SN (Figura 12) (ALLEN, 1994).

Finalizada a análise sintática que compreendeu o reconhecimento da oração por meio das regras gramaticais, a próxima etapa do PLN, denominada de análise semântica, consiste em reconhecer o significado da frase.

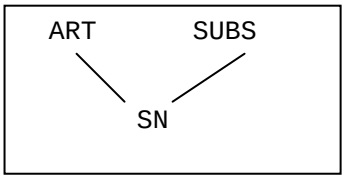

Figura 12. Derivação *BottomUp*

[10] Possui como núcleo um nome, um pronome ou substantivo. A diferença entre um nome próprio e um substantivo é que o primeiro aparece sem determinante e o segundo com determinante (SOUZA; SILVA, 2001);

[11] Apresenta como núcleo um verbo ou locuções verbais. Este verbo pode ter ou não um complemento sendo este um sintagma nominal (SOUZA; SILVA, 2001).

3.1.3 Análise Semântica

Após o reconhecimento de cada palavra e a verificação da sua posição em relação à frase, precisa-se compreender o seu significado e verificar se possui um sentido correto (HUBNER, 1992).

Posteriormente à análise sintática, as frases podem conter as suas palavras na posição correta de acordo com a árvore de derivação, porém no que se refere ao sentido semântico-gramatical podem estar erradas, como por exemplo: *O computador está andando*. Nesta frase as palavras estão na ordem correta conforme a língua portuguesa, pois possui artigo, substantivo e verbos, porém não apresenta sentido já que nenhum computador pode andar. Enquanto isso, a frase *O computador está processando* demonstra um sentido adequado ao sujeito desta frase (SCHIPIURA, 2004).

Na análise semântica pode ocorrer ambigüidade, bastante comum na língua portuguesa, que remete ao fato de uma palavra possuir mais de um sentido na mesma frase. Por exemplo, a sentença *Luto pela vida* pode demonstrar um empenho para melhorar o cotidiano, a vida diária, comportando-se, portanto como verbo. Já a mesma frase pode referir-se a um luto em função de morte, mudando completamente o sentido desta expressão. Assim, na análise da frase tenta-se proporcionar uma interpretação correta do seu sentido, a fim de deixá-la sem ambigüidade.

Portanto, a sistemática de execução do PLN compreende a aquisição da entrada inserida pelo usuário na interface e a devolução de uma resposta condizente com o questionamento. Neste caso, precisa-se capturar o comando em linguagem natural e traduzi-lo para uma linguagem formal, compreensível pelo sistema. Assim, retiram-se possíveis ambigüidades gerando-se regras lógicas. No processamento

semântico identifica-se por meio de uma palavra chave o que o usuário pretende com a consulta (RODRIGUES, 1998).

O analisador semântico é responsável por identificar o significado da frase de entrada começando pela raiz (nó sentença) e se deslocando pela árvore. Em cada nó, interpreta recursivamente os seus filhos e relaciona os resultados num grafo conceitual[12]. Por exemplo, o interpretador semântico cria uma representação de sintagma verbal construindo recursivamente filhos do nó (verbo e sintagma nominal), combinando-os para formar uma interpretação do sintagma verbal que é enviado para o nó sentença e relacionado com a representação do sujeito (LUGER, 2004).

Na realização desse processo a recursão é interrompida nos terminais da árvore de derivação. Alguns deles, como substantivos, verbos e adjetivos, fazem com que conceitos sejam recuperados da base de conhecimento. Enquanto, os artigos não correspondem a conceitos desta base, mas qualificam outros terminais do grafo (LUGER, 2004).

A representação semântica de uma frase pode ser realizada por meio de diferentes abordagens, como as gramáticas de caso e de semântica.

3.1.3.1 Gramáticas de Caso

As gramáticas de caso abordam de forma diferente a combinação entre a interpretação sintática e semântica. As regras gramaticais são escritas para descrever regularidades sintáticas, porém não semânticas. Mas as estruturas que as regras produzem correspondem as relações semânticas, não sendo estritamente sintáticas (RICH, 1988).

[12] Um grafo conceitual é um diagrama bipartido, finito, conectado, consistindo em um conjunto rotulado de nós de conceitos, um conjunto rotulado de nós de relações conceituais e um conjunto de (diretos) vínculos conceitos e nódulos de relação (MENEZES, 2005).

Considerando-se, por exemplo, as seguintes frases: *Cristiane fala inglês* e *Carlos fala inglês,* as duas frases diferenciam-se apenas pelos sujeitos *Cristiane* e *Carlos*. Portanto, pode ser conveniente reunir as duas frases originando-se *Cristiane e Carlos falam inglês,* pois as estruturas sintáticas são idênticas.

No entanto, nas frases: *O menino está voando de alegria* e *O pássaro voa alegremente,* o que as difere é o sujeito, porém é inadequado dizer que *O menino e pássaro estão voando alegremente.*

A Figura 13 apresenta uma frase, a posterior árvore de derivação sintática gerada e o uso de uma palavra chave na análise semântica por meio da gramática de caso.

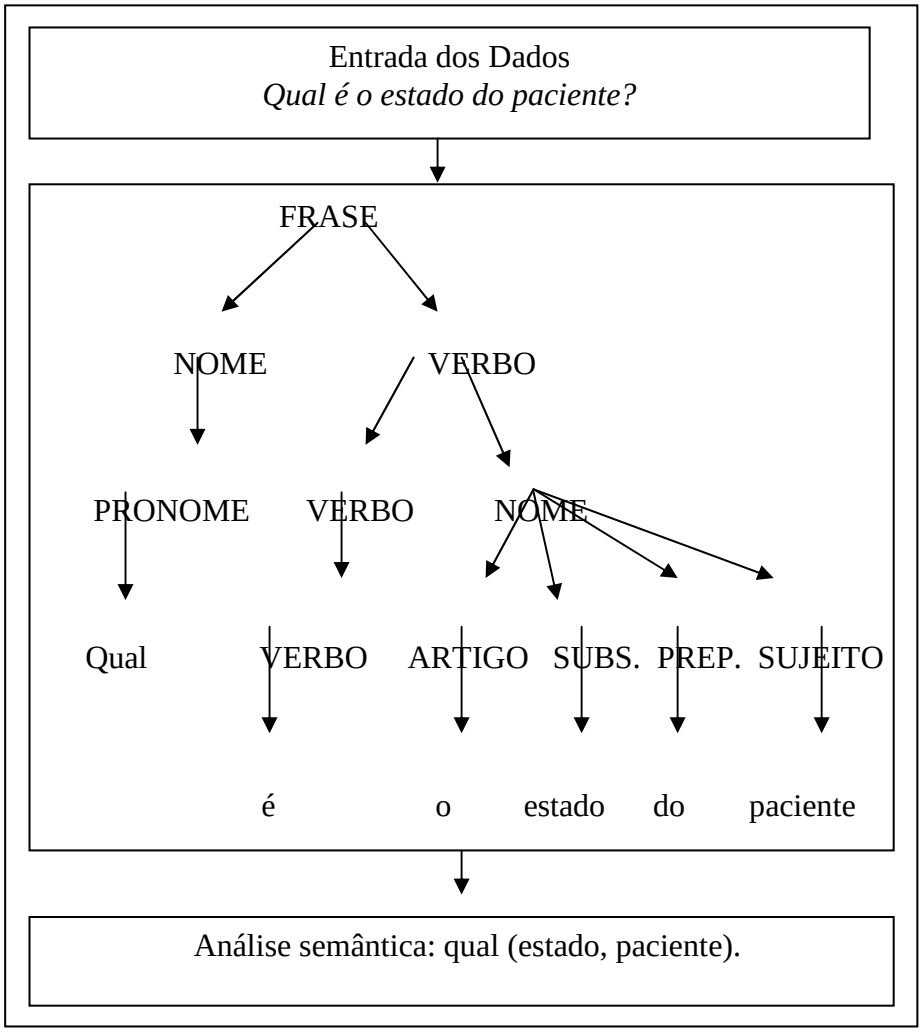

Figura 13. Derivação da Gramática de Caso

O interpretador semântico utiliza uma base de conhecimento[13] correspondente, por exemplo, a área de urologia. Os conceitos na base incluem os objetos *paciente* e *doenças* e as ações *possui* e *causa*. Segundo Luger (2004) além dos conceitos devem-se definir as relações utilizadas nos grafos conceituais, tendo-se:

a) **agente:** define a relação existente entre a ação e o objeto que a causa;

b) **experimentador:** liga um estado a um conceito;

c) **instrumento:** objeto que causa alguma ação ou utilizado para causá-la;

d) **objeto:** representa a relação existente com o verbo;

e) **parte:** liga conceitos e defini a sua relação com o todo.

A análise por meio da gramática de casos normalmente é dirigida por expectativas, assim localizando-se o verbo da frase, pode-se utilizá-lo para prever os sintagmas nominais e verbais que ocorrerão, bem como determinar o relacionamento desses com o restante da frase.

3.1.3.2 Gramáticas Semânticas

A gramática semântica é considerada do tipo livre de contexto, onde a escolha de símbolos não-terminais e das regras de produção são funções das análises semânticas e sintáticas. Resultando-se o significado da frase por meio da análise e aplicação das ações semânticas (RICH, 1988).

Este tipo de gramática não testa somente a gramaticalidade de uma sentença, mas sim produz uma análise que relata diretamente o propósito de um sistema. A ação semântica está associada a regra gramatical que é elaborada em torno de conceitos semânticos.

[13] Bases de dados ou conhecimento acumulados sobre um determinado assunto. Essas informações podem ser utilizadas na solução dos problemas apresentados pelos clientes, por meio de ferramentas de Inteligência Artificial (I.A.) ou sistemas especialistas (COELHO, 1995).

Como exemplo de gramática semântica tem-se a utilizada pelo sistema LADDER (Figura 14). Este sistema provê acesso em linguagem natural a uma grande base de dados que auxilia gerencialmente os oficiais da Marinha.

```
S → qual é PROPRIEDADE-NAVIO de NAVIO?
PROPRIEDADE-NAVIO → a PROP-NAVIO| PROP-NAVIO
PROP-NAVIO → velocidade | comprimento | contigente |
largura | tipo
NAVIO → NOME-NAVIO | o mais rápido NAVIO 2 | o mais NAVIO
2 | NAVIO 2
NOME-NAVIO → Kennedy | Constellation | ...
NAVIO 2 → PAÍSES NAVIO3 | NAVIO3
NAVIO 3 → TIPO-NAVIO LOC | TIPO-NAVIO
TIPO-NAVIO → porta-aviões | submarino | barco a remo
PAÍSES → americano | francês | inglês | russo | ...
LOC → no Mediterrâneo | no Pacífico | ...
```

Figura 14. Gramática Semântica do Sistema LADDER
Fonte: RICH, E. (1988)

3.1.4 Análise Pragmática

A composição de cada frase escrita ou falada possui um contexto tanto no momento de expressá-la quanto de entendê-la. Sendo assim, a análise pragmática se utiliza basicamente da relação existente entre os símbolos e as pessoas que deles se utilizam (COSTA; SIMÕES, 2004).

Exemplificando-se a análise pragmática, considere as seguintes frases de um trecho poético, neste contexto indica-se pelo pronome *se* que o amor foi perdido.

> Se se morre de amor!
> No mesmo tempo, e ser no mesmo ponto
> O ditoso, o misérrimo dos entes: Isso é amor, e desse amor se morre! (DIAS, 1982, p.36).

Sob o aspecto pragmático, interessam os efeitos que a linguagem produz entre os membros de uma comunidade, estudando-se as relações sociais por meio do uso concreto da linguagem (ALLEN, 1994).

Dessa forma, os elementos (tempo, lugar, quem, situação, entre outros) moldam o ambiente em que se faz um diálogo. Esses ambientes definem da forma mais clara o que o interlocutor quer expressar (FIORIN, 2003).

O interlocutor encontra-se em um determinado ambiente, onde a frase por ele expressada é formada em função do pronome *eu* que se torna o sujeito e ponto de referência. A seguir, tem-se um exemplo que descreve a presença deste pronome pessoal.

Encontrei com Pedro, que me disse:

- *Estou muito insatisfeito com a minha relação na empresa.*

Neste diálogo verifica-se a presença de alguns *eu* sendo todos eles diferentes, no caso encontrei-me remete ao *eu* primeira pessoa, já no caso de estou o *eu* refere-se à terceira pessoa, no caso o Pedro (FIORIN, 2003).

Dessa forma, conclui-se que todo o discurso demonstra como é a relação entre o sujeito e o predicado da oração, denotando algumas vezes o sentido da frase. Portanto, explorar um texto compreende a identificação dos elementos que o compõe e as maneiras como eles interagem. Assim, podem-se obter as suas características específicas e determinar o contexto a que ele pertence.

Segundo Rich (1988) a análise pragmática ocorre por meio do relacionamento existente entre as frases, necessitando-se de conhecimento significativo para o reconhecimento destas relações. Para isso, pode-se utilizar:

a) **foco na compreensão:** o uso do conhecimento para auxiliar na compreensão deve ser voltado à parte relevante de uma base de conhecimento a fim de resolver ambigüidades e relacionar o que foi informado. Considerando-se o seguinte texto: *Realize o procedimento cirúrgico no paciente. Os instrumentos encontram-se esterilizados*.

Necessita-se para compreender este texto reconhecer que os *instrumentos* citados na segunda frase são utilizados para *realizar o procedimento*. Isto pode ser realizado, se no momento da compreensão da primeira frase os *instrumentos* forem trazidos a foco. Portanto, precisa-se de uma forma para representar o conhecimento a respeito de *realizar*, de forma que sempre ao se citar uma *realização*, conceitos associados, como o de que *instrumentos* são necessários, sejam acessados;

b) **estruturas de meta para a compreensão:** consiste na identificação dos personagens e seus planos para atingir determinadas partes de uma meta. Por exemplo: *Marcos necessitava de uma consulta médica. Ele perguntou ao homem que passava por ele onde ficava o Posto de Saúde mais próximo.* Nesta expressão tem-se a meta (realizar a consulta médica) e o plano (ir ao Posto de Saúde). Este plano exige o conhecimento de onde há um posto de saúde, fixando-se portanto, outra meta (saber onde se encontra este posto mais próximo). Assim, o plano de Marcos para atender esta submeta consiste em perguntar para alguém;

c) **esquemas e roteiros:** descrições detalhadas de padrões específicos de ocorrência comum são armazenadas. Assim, empregam-se os roteiros para auxiliar na compreensão da linguagem natural. Os roteiros contêm informações específicas sobre uma determinada situação e permitem a realização da inferência. Dessa forma, consistem em uma estrutura mais dinâmica que se relaciona em determinados graus de generalidade e descreve experiências, como por exemplo, a de ir a uma consulta médica;

d) **compreensão de diálogo:** meio pelo quais as pessoas transmitem dados, tornando-se uma aplicação importante para a comunicação entre o

> computador e seus usuários. Neste processo de compreensão se deve explorar todas as fontes de informações disponíveis referente ao diálogo, incluindo: a sua característica a ser acionada pela meta; o seu foco atual; as regras utilizadas, entre outros.

A estrutura do PLN composta pelas análises léxica, sintática, semântica e pragmática foi abordada nesta seção. A seguir, destacam-se algumas das aplicações desta técnica da inteligência artificial.

3.2 APLICAÇÕES DO PROCESSAMENTO DE LINGUAGEM NATURAL

O processamento de linguagem natural tem sido aplicado no desenvolvimento de ferramentas computacionais que possuem diferentes funcionalidades, como por exemplo, para: tradução; escrita e produção de texto; sistemas interativos; extração de informações e interfaces.

A primeira área que se aplicou o PLN foi para a tradução de documentos, porém essa atividade exige muito conhecimento lingüístico já que codifica a informação escrita em um determinado idioma para outro (LUGER, 2004).

Na redação e produção de texto têm-se programas que identificam erros, sugerem alternativas, disponibilizam recursos como dicionários de dados (monolingues, bilíngües, sinônimos, entre outros) e proporcionam auxílio no que se refere a gramática. Nesta área encontram-se também aplicações que produzem semi-automaticamente documentos especializados. Para isso, parte-se de informações presentes em bases de dados e auxilia-se na redação de boletins metereológicos, criação de páginas *web*, entre outros (BOURBEAU et al, 1990; DIMARCO; FOSTER, 1997). Além de ferramentas deste tipo, têm-se aquelas que convertem um discurso oral em texto escrito.

Nos sistemas interativos pode-se, embora de forma ainda primária, fornecer comandos a uma máquina por meio da fala, como por exemplo, em carros, casas inteligentes, organizadores pessoais, entre outros. Portanto, deve-se salientar que esta abordagem possui potencial para aplicação em jogos de computador a fim de auxiliar na interação com os personagens (SANTOS, 2001).

A extração de informações por meio de PLN pode ser aplicada a documentos na internet a fim de localizar a informação sobre um domínio especifico e codificá-la numa forma adequada para relatar ao usuário ou até mesmo para construir um banco de dados estruturado (LUGER, 2004).

O PLN também pode ser utilizado para o desenvolvimento de interfaces em sistemas computacionais, aceitando consultas em linguagem natural e fornecendo respostas de forma escrita. Esta área constitui-se em uma aplicação importante desta tecnologia, podendo ser empregada para auxiliar na interação homem-computador (LUGER, 2004).

Considerando-se que esta pesquisa compreendeu o desenvolvimento de uma interface em PLN para a área médica, a seguir aborda-se mais detalhadamente a aplicação de linguagem natural em interfaces de sistemas computacionais.

3.2.1 Interface em Linguagem Natural

A interface com o usuário é o mecanismo pelo qual se pode estabelecer uma interação entre o programa e o ser humano. Por exemplo: a tela de um determinado software, bem como um ícone na área de trabalho referem-se a interface entre o usuário e o sistema. Portanto, pode-se concluir que uma interface consiste em um sistema de comunicação (PRESSMAN, 2001).

A Figura 15 mostra o tipo de interface comumente apresentada nos softwares, sendo composta por menus, caixas de textos, botões, entre outros componentes gráficos que visam auxiliar na interação e facilitar a utilização dos sistemas pelos usuários. A reunião desses elementos, bem como o uso de cores, formas e o acesso às funcionalidades do sistema por meio do mouse e teclado podem também auxiliar os indivíduos.

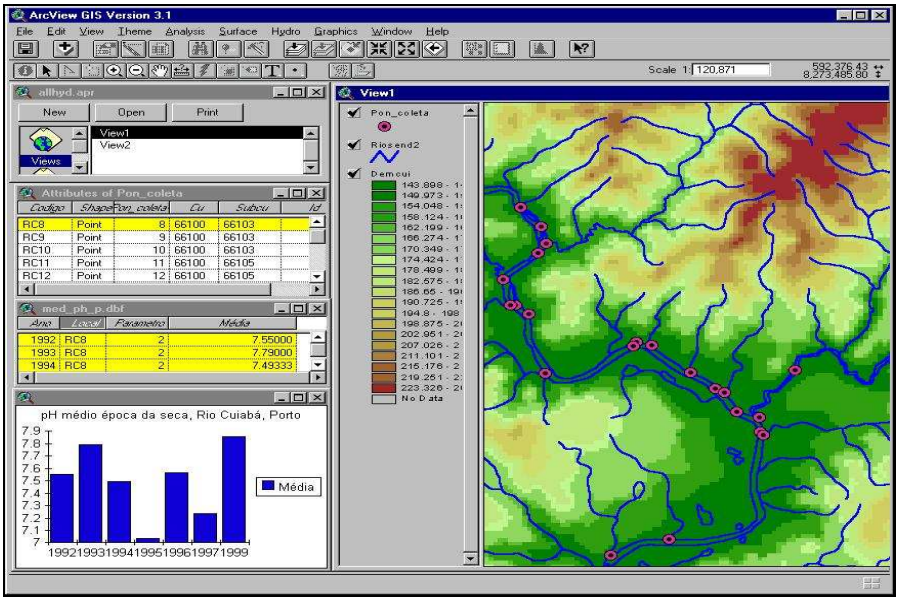

Figura 15. Exemplo de Interface Gráfica
Fonte: UFMT (2000)

De acordo com Pressman (2001) os componentes gráficos de uma interface devem estar ligados há algo que procure mimetizar objetos familiares, como por exemplo, disquetes, folhas de papel, impressoras, entre outros (Figura 16). Assim, remete-se o usuário àquele componente (imagem) que representa a tarefa que ele pretende executar.

Figura 16. Ícones

O desenvolvimento tecnológico tem disponibilizado aos usuários diferentes componentes de hardware e de software, bem como interfaces que auxiliem no processo de interação ser humano – computador. Mediante isso, têm-se diferentes opções com

esse fim, porém muitas vezes tornam-se inviáveis por confundirem o usuário e demandarem de conhecimento prévio, bem como de tempo para aprendizagem (HÜBNER, 1992).

Contrapondo-se a este conceito, a linguagem natural possui uma capacidade de expressão superior, portanto, utilizar a linguagem natural na interface de comunicação consiste na troca de informações entre o usuário e o computador de maneira em que os dois se entendam. Dessa forma, o usuário não precisa se adaptar com a estrutura computacional (SCHIPIURA, 2004).

Na interface em linguagem natural (Figura 17) usa-se basicamente a conversação como meio de interação entre usuário e máquina. Retirando-se, assim, muitas opções de menus utilizadas para interação com um software qualquer. No que se refere as interfaces em linguagem natural tem-se a utilização de um campo para digitação de perguntas e outro onde o programa disponibilizará ao usuário a resposta de forma escrita em linguagem natural. Assim, utilizando-se frases para a interação do usuário com a interface, cria-se um diálogo, que é uma característica da linguagem natural nas interfaces de computadores.

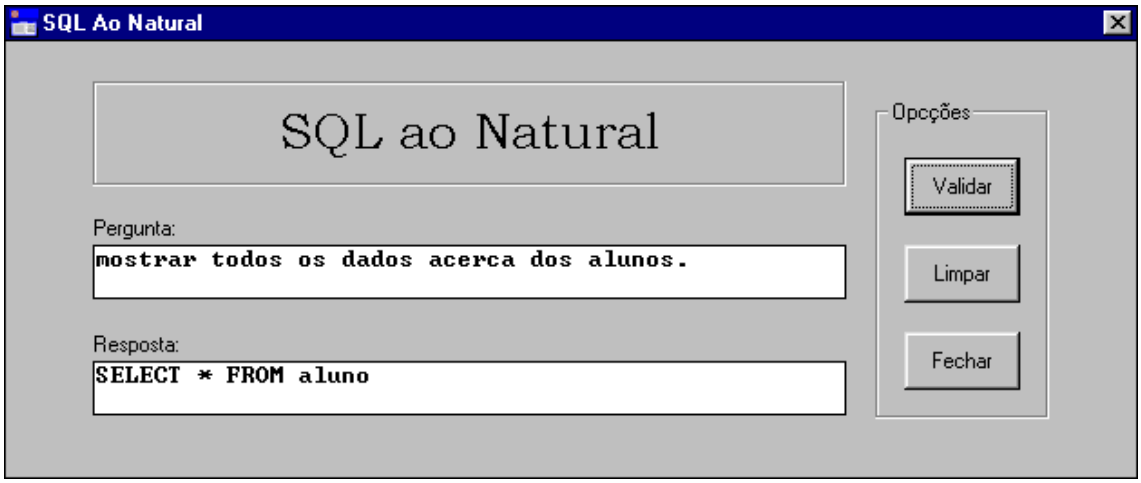

Figura 17. Interface PLN para Banco de Dados

Este tipo de interface computacional pode ser aplicado em vários tipos de sistemas como de apoio à informática na saúde que possui um rápido desenvolvimento

científico e trabalha com armazenamento, recuperação e uso da informação, dados e conhecimentos biomédicos para a resolução de problemas e tomada de decisão.

4 INFORMAÇÃO E INFORMÁTICA EM SAÚDE

A informática em saúde consiste no conceito ampliado que engloba os dados fornecidos por várias instituições públicas ou privadas de atenção à saúde. As informações em saúde oferecem subsídios para programas de ações e tomada de decisão, auxiliando na promoção da saúde e prevenção da doença. Na realização da mensuração desses dados são considerados alguns fatores, dentre os quais: condição de vida, fatores ambientais, qualidade e acesso aos serviços de saúde, entre outros. Esses fatores são essenciais para a construção dos dados a serem posteriormente utilizados na composição de informações para a área da saúde (DATASUS, 2006).

A informação em saúde compreende todas aquelas que se relacionam as doenças, condições de morte e ao nascimento de indivíduos e populações bem como a gestão administrativa. Estas informações podem ser úteis para estipular diretrizes políticas na área, como também para identificar condições sócio-ecônomicas e desigualdades sociais presentes. Além disso, a informação em saúde está relacionada entre com as tecnologias que auxiliam à sua produção e disseminação, o que contribui para a adoção de prontuários eletrônicos nos hospitais, consolidação da Rede Nacional de Informação em Saúde[14] e do Cartão Nacional de Saúde[15] (MORAES, 2002).

Portanto, quando se discute a informação em saúde é preciso inserir as questões de tecnologia da informação. Conforme a Associação Brasileira de Saúde Coletiva (2003) quando se trabalha com informação em saúde devem-se aprimorar os

[14] Um projeto de uma rede integrada, na INTERNET, para prover acesso e intercâmbio de informações em Saúde para gestão, planejamento e pesquisa para gestores, agentes e usuários do SUS (DATASUS, 2007)

[15] Instrumento que possibilita a vinculação dos procedimentos executados no âmbito do Sistema Único de Saúde (SUS) ao usuário, ao profissional que os realizou e também à unidade de saúde onde foram realizados (DATASUS, 2007)

sistemas e metodologias de produção e disseminação destas informações, bem como a utilização de tecnologias voltadas a construção de padrões[16].

Atualmente, tem-se preocupado com a utilização desses padrões para a representação da informação e conhecimento na área da saúde, no que se refere a vocabulários e conteúdos, o que é de fundamental importância para a troca de informações, construção do prontuário eletrônico e descoberta de novas relações e conhecimentos que somente poderão ocorrer à medida que metodologias e padrões estiverem definidos.

A informação de um modo geral, bem como na área da saúde, é um produto que possui a capacidade de gerar conhecimento. Este por sua vez é composto por valores e crenças que determinam o que se aprende e conclui por meio das observações. Porém, além da produção do conhecimento são necessários meios capazes de o reunir, organizar e gerir.

Assim, sistemas voltados para o gerenciamento de informações em saúde devem possuir tanto dados médicos quanto administrativos. Isto se deve ao fato das informações estarem sempre presentes no local de atendimento ou disponíveis para a realização de consultas rápidas e precisas, gerando-se bases de dados aplicadas as diferentes áreas da saúde, como por exemplo, odontológica, farmacêutica, médica, entre outras.

As informações em saúde são disponibilizadas a partir de aplicativos como prontuários eletrônicos gerando bases de conhecimentos.

[16] A utilização de padrões permitirá a utilização eficiente de sistemas de informação em saúde possibilitando a realização de pesquisas e busca de informações bem como a integração com outros sistemas informatizados (DATASUS< 2007)

4.1 PRONTUÁRIO ELETRÔNICO DO PACIENTE

A Resolução CFM nº 1.638/2002 artigo 1º define prontuário eletrônico como:

> consistem em um conjunto de informações, sinais e imagens registradas, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a assistência a ele prestada, de caráter legal, sigiloso e científico, que possibilita a comunicação entre membros da equipe multiprofissional e a continuidade da assistência prestada ao indivíduo.

De modo geral, o princípio básico de construção do PEP baseia-se na integração da informação clínica e administrativa de pacientes individualmente. Assim, uma vez coletada a informação, ela é registrada em um determinado formato para fins de armazenamento e tal registro passa a ser fisicamente distribuído entre os hospitais, agências de seguro-saúde, clínicas, laboratórios e demais setores envolvidos, sendo compartilhado entre os profissionais de saúde, de acordo com os direitos de acesso de cada um. Além de integração, um dos requisitos básicos do PEP é a interoperabilidade, que é a habilidade de dois ou mais sistemas computacionais trocarem informações, de modo que a informação trocada possa ser utilizada (USP, 2003).

O primeiro passo para desenvolver um PEP é o entendimento de que a construção do prontuário eletrônico é um processo. Primeiro todas as informações são geradas por computadores, depois requer que o sistema esteja implantado em toda a instituição contenha elementos como integração com sistema de gerenciamento da prática, sistemas especialistas como alertas clínicos e programas de educação ao paciente. Em um nível mais elevado da informação, esta não é baseada somente nas necessidades do serviço de saúde; é baseada na saúde e doença do individuo da comunidade sendo a informação do paciente é única e feita a nível nacional.

A reunião das informações na área da saúde podem ser padronizadas para que possam ser utilizadas e entendível por qualquer usuário.

4.2 PADRONIZAÇÃO DA INFORMAÇÃO EM SAÚDE

A padronização no armazenamento e no uso do vocabulário médico é fundamental para reunir informações clínicas no cuidado ao paciente e no tratamento da doença, auxiliando na pesquisa e na condução dos resultados médicos (ALEXANDRINI, 2005).

Existem alguns padrões que proporcionam esta padronização do vocabulário, como por exemplo, (DATASUS, 2007):

a) **Tabela SUS:** pretende estabelecer um padrão para informações que são descritas em prontuário eletrônicos, definindo a participação do paciente, do conteúdo e toda a estrutura lógica do prontuário. Esta é uma maneira encontrada para padronizar as informações captadas em todo o país;

b) ***Digital Imaging Communications in Medicine*** **(DICOM):** é uma série de regras que permite que imagens médicas e informações associadas sejam trocadas entre equipamentos de imagem, computadores e hospitais. O padrão estabelece uma linguagem comum entre os equipamentos de marcas diferentes, que geralmente não são compatíveis, e entre equipamentos de imagem e computadores, estejam esses em hospitais, clínicas ou laboratórios;

c) ***Unified Medical Language System*** **(UMLS):** compreende muitos vocabulários para a área de ciências biomédicas. Provê uma estrutura de

modelagem entre estes vocabulários e assim permite traduzir entre os vários sistemas de terminologia;

d) ***Systematized Nomenclature of Medicine*** **(SNOMED):** nomenclatura criada para indexar o conjunto de registros médicos que inclui todos os tipos de sinais e sintomas, diagnósticos e procedimentos. Em um diagnóstico na SNOMED, este possui os códigos: topográfico (anatômico), morfológico (alterações em células), de organismo vivo (organismo com vida) e funcional (sinais e sintomas);

e) **Classificação Internacional de Doenças (CID):** esta padronização foi criada pela Organização Mundial de Saúde (OMS) com o propósito de classificar e padronizar as informações médicas.

A padronização da informação da base de dados utilizada pelo Hades corresponde a CID-10 versão 10, que possui como padronização normas da Organização Mundial de Saúde.

4.1.1 Classificação Internacional de Doenças Versão 10

A Classificação Internacional de Doenças versão 10 (CID-10) é a última revisão da CID que iniciou em 1983, utilizada a partir de 1996 contando com o apoio de todos os países que compõem a OMS, os quais encaminharam propostas para esta última revisão. Além destes países, grandes grupos de especialistas da área ajudaram na composição da CID-10 (DATASUS, 2007).

Foram criadas novas inovações para a CID-10 em relação à versão 9, dentre elas a principal foi o uso de um esquema de código alfanumérico que consiste em uma letra seguida de três números. Considerando-se todas as letras do alfabeto, apenas a U

não foi utilizada, a fim de servir para adições futuras. A CID-10 apresenta vinte e um capítulos (Tabela 1).

Tabela 1. CID-10

Capítulos	Blocos de Códigos	Títulos
I	A00-B99	Determinadas doenças infecciosas e parasitarias
II	C00-D48	Neoplasias
III	D50-D89	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
IV	E00-E90	Doenças endócrinas, nutricionais e metabólicas
V	F00-F99	Transtornos mentais e comportamentais
VI	G00-G99	Doenças do sistema nervoso
VII	H00-H59	Doenças do olho e anexos
VIII	H60-H95	Doenças do ouvido e da apófise mastóide
IX	I00-I99	Doenças do aparelho circulatório
X	J00-J99	Doenças do aparelho respiratório
XI	K00-K93	Doenças do aparelho digestivo
XII	L00-L99	Doenças da pele e do tecido subcutâneo
XIII	M00-M99	Doenças do sistema osteomuscular e do tecido conjuntivo
XIV	N00-N99	Doenças do aparelho geniturinário
XV	O00-O99	Gravidez, parto e puerpério
XVI	P00-P96	Algumas afecções originadas no período perinatal
XVII	Q00-Q99	Malformações congênitas, deformidades e anomalias cromossômicas.
XVIII	R00-R99	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte.
XIX	S00-T98	Lesões, envenenamento e algumas outras conseqüências de causas externas.
XX	V01-Y98	Causas externas de morbidade e de mortalidade
XXI	Z00-Z99	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde

Fonte: DATASUS (2007)

Nesta classificação o Sarampo, por exemplo, apresenta como código B05. Portanto, esta morbidade encontra-se no capítulo 1 que se refere as doenças infecciosas e parasitárias. Cada letra do código identifica o capítulo e os números designam a especificação para cada morbidade (variando de 0 a 100).

A partir dos conhecimentos apresentados de PLN (Capítulo 3) e de informações em saúde (Capítulo 4), que compreendem a base desta pesquisa, têm-se alguns exemplos de trabalhos correlatos sobre a utilização de PLN na área da saúde, em específico para apoio a prática médica.

5 TRABALHOS CORRELATOS

A utilização da técnica de processamento de linguagem natural tem se difundido, sendo este um dos objetivos desde o princípio da inteligência artificial. Porém, ainda são poucas as ferramentas desenvolvidas para a língua portuguesa, constituindo-se, portanto, em um campo de pesquisa.

Dentre as publicações na área a maioria constitui-se em bibliografia estrangeira, sendo que alguns desses estudos de PLN estão cada vez mais sendo inseridos em aplicações na área médica o que auxilia no desenvolvimento de ferramentas que possam beneficiar estes profissionais.

5.1 DESENVOLVIMENTO DE UMA METODOLOGIA DE INTERPRETAÇÃO, RECUPERAÇÃO E CODIFICAÇÃO INTELIGENTE DE LAUDOS MÉDICOS INDEPENDENTE DE IDIOMA

Alexandrini (2005) desenvolveu uma Tese de Doutorado na Universidade Federal de Santa Catarina sobre metodologias inteligentes de laudos médicos. O trabalho centra-se no uso de métodos de interpretação, codificação e recuperação inteligentes em laudos médicos utilizando a técnica de PLN combinada com terminologias internacionais. A tese descreve uma ferramenta (Figura 18) que faz a recuperação e interpretação de laudos médicos em padrão texto.

Figura 18. Interface Principal do ACLM
Fonte: ALEXANDRINI, F. (2005)

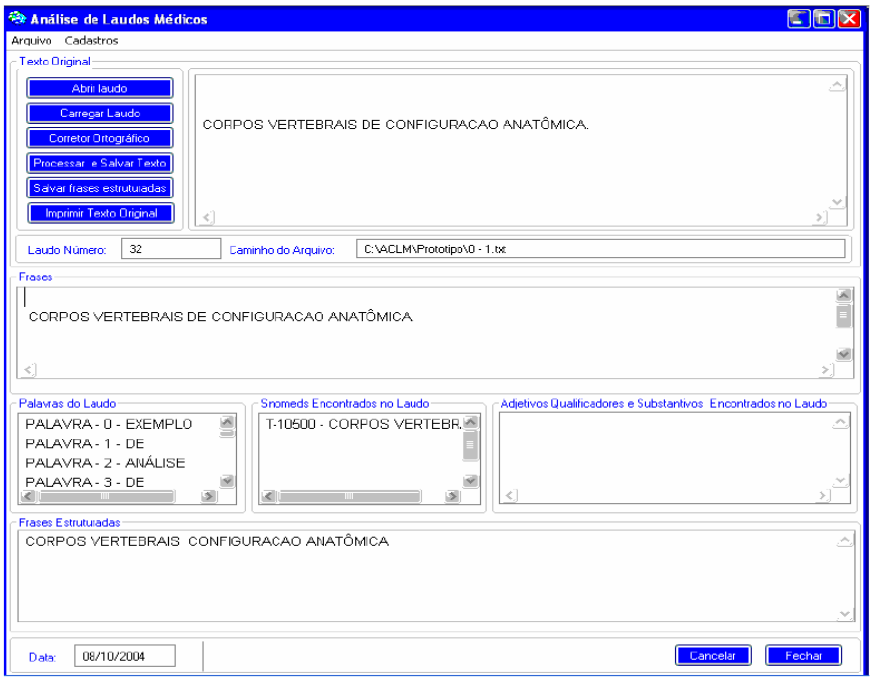

Figura 19. Interface de Análise de Laudos
Fonte: ALEXANDRINI, F. (2005)

Esta ferramenta utiliza a SNOMED como norma para a estruturação dos laudos e visa a integração com softwares de edição baseados no padrão DICOM SR-Struct Report[17].

[17] Abreviação de Digital Imaging Communications in Medicine, é conjunto de normas para tratamento, armazenamento e transmissão de informação médica (imagens médicas) num formato eletrônico. (ALEXANDRINI, 2005).

5.2 CAS: UMA INTERFACE EM LINGUAGEM NATURAL, UTILIZANDO A MEMÓRIA DINÂMICA E RACIOCÍNIO BASEADO EM CASO PARA AUXÍLIO NA GERAÇÃO DE DIAGNÓSTICOS

Schipiura (2004) elaborou uma Dissertação de Mestrado pela Pontifica Universidade Católica do Paraná na área de Informática na Saúde que compreendeu o desenvolvimento de uma interface em linguagem natural para auxílio à decisão médica. Esta interface possibilita aos profissionais da área médica elaborar questões ao sistema e obter respostas nos mesmos moldes em que o questionamento foi realizado. Esta dissertação teve como objetivo um sistema que compreende o significado dos termos e das expressões em saúde que os profissionais utilizam, gerando informações nos mesmos padrões.

O software desenvolvido é composto por quatro módulos sendo eles, os geradores de memória, questões, respostas e o analisador dos resultados. O profissional de saúde ao utilizar o sistema proporciona que este incorpore o conhecimento relativo a sua área de domínio. Para isso, usa-se uma memória dinâmica que a partir das informações em uma base de dados gera uma base de conhecimento (SCHIPIURA, 2004). Na interpretação das questões e respostas (Figura 20) utiliza-se um *parser*.

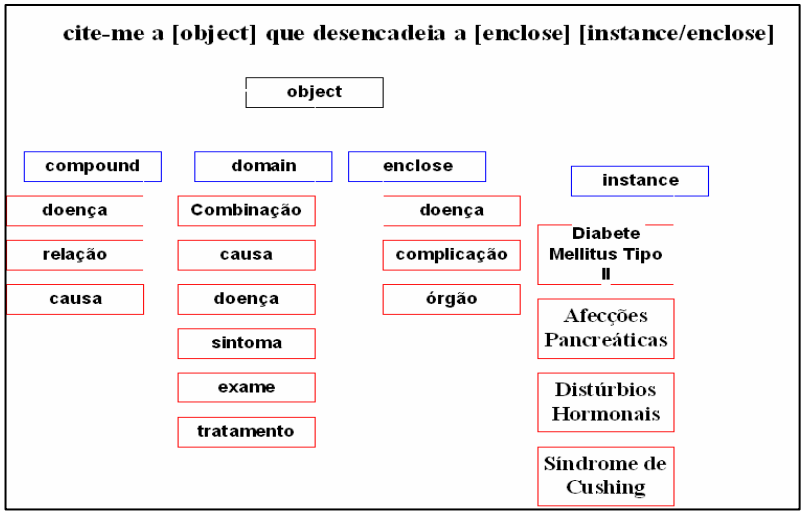

Figura 20. Parser Reconhecimento Sentença
Fonte: SCHIPIURA, C. (2004)

5.3 UM MODELO DE HIPERDICIONÁRIO: ESTUDO DE CASO EM PRONTUÁRIOS MÉDICOS

Wives (1996) realizou essa pesquisa como Trabalho de Conclusão de Curso na Universidade Católica de Pelotas. Apresenta um modelo de ferramenta que procura solucionar um dos problemas da busca de informações textuais, recuperando documentos realmente relevantes ao usuário, por meio da análise de contexto.

No desenvolvimento utilizou uma estrutura de hiperdicionário capaz de armazenar palavras e as relações entre estas. Possibilitando, portanto, a recuperação de informações textuais não simplesmente por palavras-chave, mas também por um conjunto de palavras inter-relacionadas que caracterizam determinado contexto. Estes contextos foram utilizados em uma clínica local a fim de recuperar informações e avaliar este hiperdicionário (WIVES, 1996).

5.4 PROCESSAMENTO DE LINGUAGEM NATURAL E A REPRESENTAÇÃO DE DADOS CLÍNICOS

Saccer et al (1994) desenvolveram na Universidade de Nova Iorque uma forma de representar observações e ações clínicas por meio de um método de processamento de texto livre.

Esse projeto emprega a análise sintática por meio de uma gramática de sublinguagem e informação estruturada que são especificadas para a narrativa clínica. No desenvolvimento desse trabalho foram elaboradas perguntas por meio da medição da qualidade sendo feita por meio de informações precisas e recuperáveis, para isso foram elaborados perguntas sobre 13 critérios de asma em uma base de dados.

5.5 ESTUDO DE UMA INTERFACE DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA REGISTRO MÉDICO

Takemura e Ashida (2002) da Universidade de Osaka no Japão estudaram como a informação sobre um paciente pode ser acessada em um sistema computadorizado que utilize a linguagem natural. Portanto, desenvolveram um protótipo de sistema com uma interface que proporciona a interação entre computador e homem por meio de PLN.

Considerando-se os exemplos de trabalhos correlatos, apresentados neste capítulo, realizou-se o desenvolvimento desta pesquisa por meio da implementação de um sistema de consulta em linguagem natural, denominado Hades, a uma base de dados da área de urologia.

6 INTERFACE EM PROCESSAMENTO DE LINGUAGEM NATURAL PARA UROLOGIA

A pesquisa aqui apresentada consistiu no desenvolvimento de uma interface em processamento de linguagem natural denominada de Hades. Esta aplicação por meio de seus métodos possui como finalidade auxiliar estudantes e profissionais da área de urologia, possibilitando a realização de consultas a uma base de dados e a extração de informações necessárias.

A implementação da interface compreendeu quatro módulos de análises que realizam o PLN em interfaces naturais. Para isso, realizou-se um estudo referente a inteligência artificial, PLN, interface em linguagem natural, dentre outros.

Finalizada a implementação do Hades, realizaram-se testes por meio de uma base de dados que apresenta informações referentes a algumas consultas médicas da área de Urologia.

6.1 BASE DE DADOS UTILIZADA

A fim de demonstrar o funcionamento da interface em PLN utilizou-se uma base de dados referente à pacientes de uma Clínica de Urologia no Município de Criciúma - SC. Esta base composta por dados pessoais, clínicos, dentre outros, foi disponibilizada pelo especialista do domínio de aplicação, o médico urologista Rozenir Ramos, professor do Curso de Medicina da Universidade do Extremo Sul Catarinense.

A Urologia é uma especialidade clínica e cirúrgica responsável pelos sistemas urinários de ambos os sexos e do aparelho reprodutor masculino. O sistema

urinário é formado pelos rins, ureteres, bexiga e uretra (Figura 21) (BARATA; CARVALHAL, 1999).

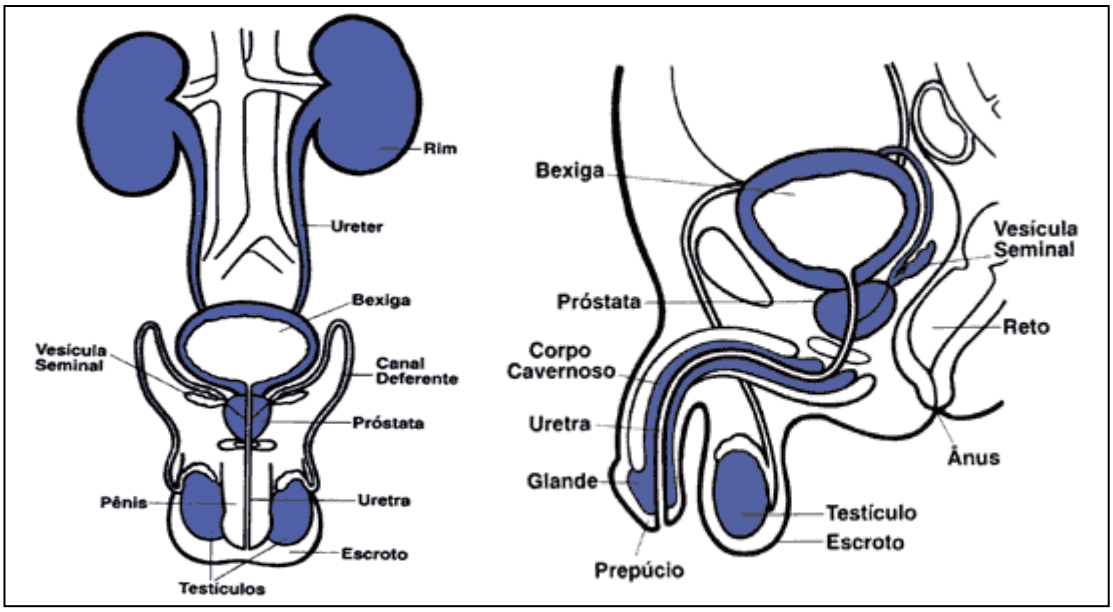

Figura 21. Aparelho Urinário e Reprodutor
Fonte: SBU (2007)

As principais doenças de que trata a urologia são (SBU, 2007):

a) **câncer de próstata**: a próstata é uma glândula localizada próxima a bexiga, bastante suscetível ao desenvolvimento de câncer em função de fatores genéticos e hormonais, sendo mais comum nos homens que possuem idade superior a 50 anos;

b) **hiperplasia prostática benigna**: caracteriza-se pelo aumento da próstata, não estando associada ao câncer. Ela pode ser causada em função da idade (acima dos 50), fatores genéticos e hormonais;

c) **cálculo renal**: são cristais que se separam da urina e se unem formando pedras, sendo mais comum em homens. As causas podem estar relacionadas a fatores genéticos, infecções urinárias e distúrbios renais;

d) **fimose**: é a incapacidade de expor por completo a glande;

e) **trauma peniano**: pode ocorrer durante a relação sexual quando ocorre a ruptura dos corpos cavernosos, bem como em casos de queimaduras ou cortes no pênis.

6.2 METODOLOGIA

O desenvolvimento do Hades fundamentou-se metodologicamente pelas seguintes etapas: levantamento bibliográfico; definição da estrutura do Hades; modelagem da interface; implementação das análises léxica, sintática, semântica e pragmática no desenvolvimento de uma ferramenta inteligente para consultas, finalizando-se com a realização de testes.

O levantamento bibliográfico compreendeu a pesquisa dos temas envolvidos neste trabalho de forma a se entender e descrever a linguagem, o PLN, a informação em saúde, entre outros. Durante a pesquisa, os estudos foram dedicados principalmente ao entendimento das análises envolvidas durante a execução do PLN, já que esta constitui a base desta pesquisa.

6.2.1 Definição da Estrutura do Hades

Estruturalmente o Hades é composto por cinco módulos: interface de entrada dos dados; análise léxica, sintática, semântica e pragmática, e finalmente o resultado da consulta para o usuário (Figura 22).

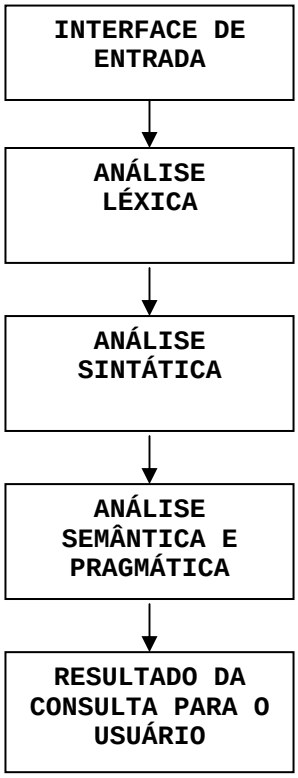

Figura 22. Fluxograma da Estrutura

6.2.2 Modelagem do Protótipo Hades

Na modelagem do Hades utilizou-se a *Unified Modeling Language*[18] (UML), realizando-se os diagramas de caso de uso, atividades e seqüência por meio da ferramenta JUDE que possui uma versão gratuita disponível no endereço http://jude.change-vision.com/jude-web/index.html.

O diagrama de caso de uso demonstra as atividades a serem realizadas pelo usuário na interface, bem como a função do sistema (Figura 23):

[18] Uma linguagem para especificação, documentação, visualização e desenvolvimento de sistemas orientados a objetos. Sintetiza os principais métodos existentes, sendo considerada uma das linguagens mais expressivas para modelagem de sistemas orientados a objetos (FOWLER; SCOTT, 2000).

a) **entrada dos dados**: o usuário informa a pergunta que deseja consultar na base de dados, além disso, solicita ao sistema a execução dos algoritmos que realizam as análises sobre a pergunta;

b) **execução dos algoritmos:** uma vez realizada a solicitação pelo usuário o sistema executa os processos de análises envolvidos para o PLN no Hades.

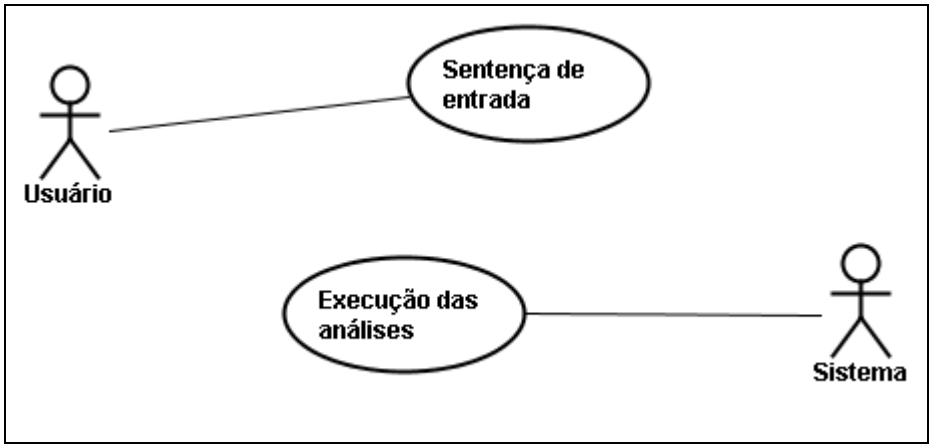

Figura 23. Diagrama de Caso de Uso

A Figura 24 demonstra o diagrama de atividades que possibilita a identificação do fluxo de tarefas executadas pelo usuário e o sistema:

a) **informar dados de entrada:** o usuário informa a pergunta necessária para a execução do sistema;

b) **solicita execução do algoritmo:** por meio de um botão o usuário solicita a execução dos algoritmos utilizados nas análises da linguagem;

c) **execução do algoritmo:** execução do sistema utilizando os algoritmos para as análises léxicas, sintáticas, semânticas e pragmáticas;

d) **resultados:** após a execução do algoritmo o resultado é mostrado ao usuário com base nos dados da área de Urologia.

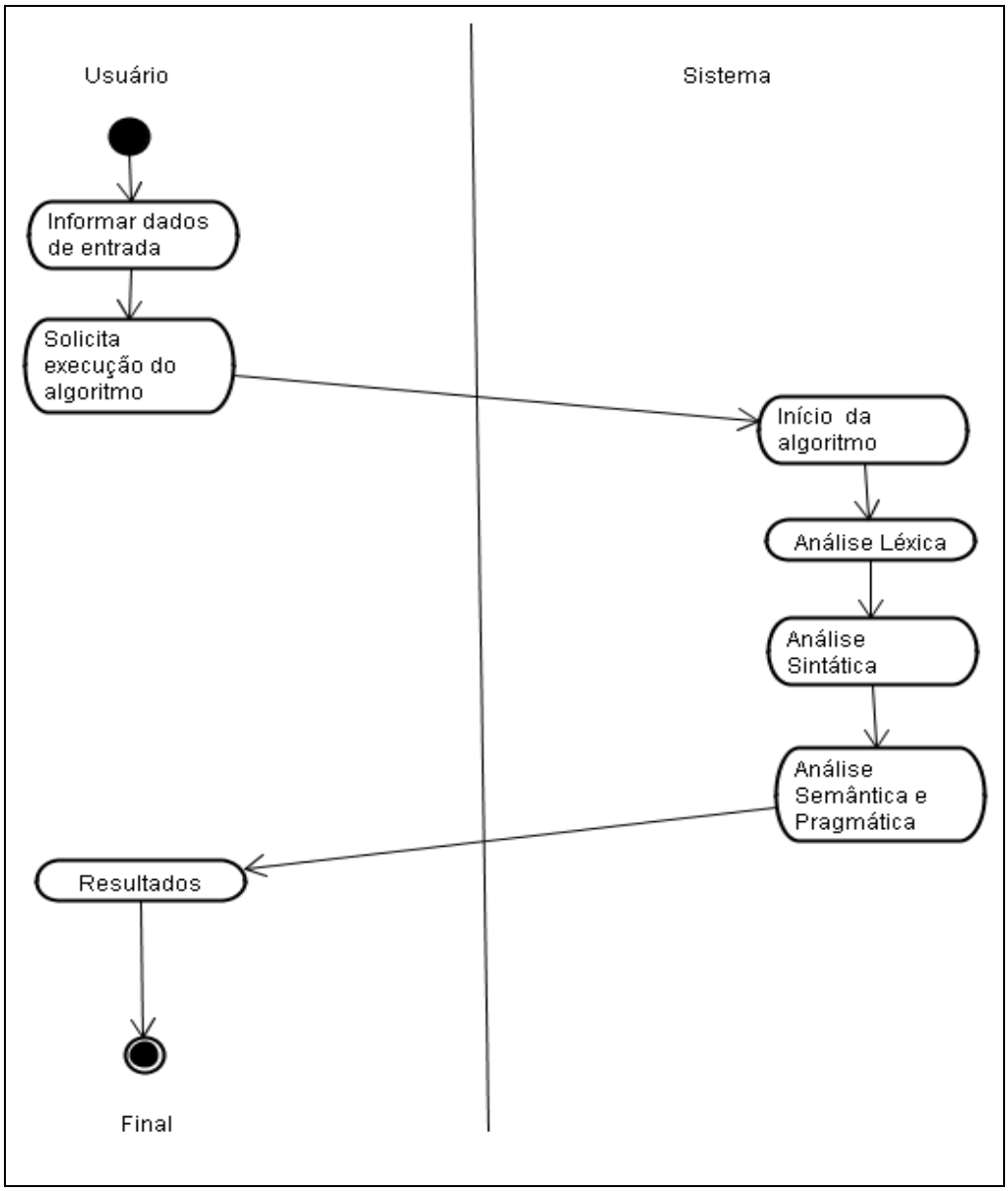

Figura 24. Diagrama de Atividades

O diagrama de seqüência apresenta o processo onde o usuário solicita o método de PLN e após a sua execução os resultados retornam a ele. O método de PLN consiste em executar os módulos de análises e busca dos dados na base bem como no retorno do resultado para a tela do sistema (Figura 25).

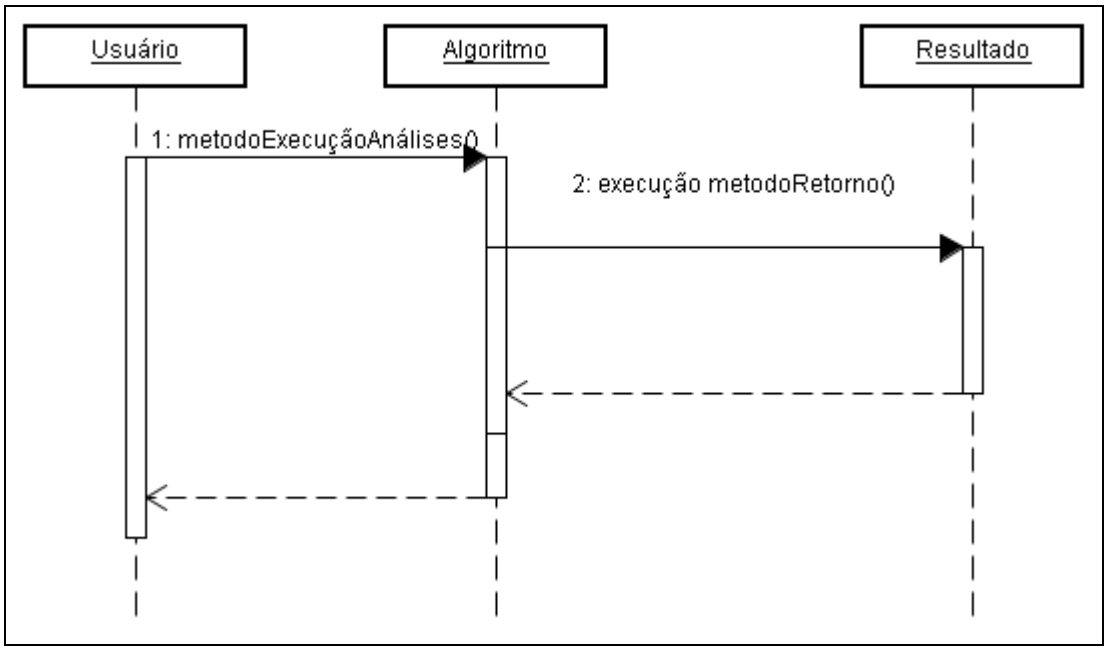

Figura 25. Diagrama de Seqüência

6.2.3 Implementação da Análise Léxica

Esta primeira etapa do PLN no Hades refere-se ao reconhecimento da frase palavra por palavra (*tokens*) buscando-as no dicionário criado pelo sistema, utilizando para isso cinco classes gramaticais.

As classes gramaticais adotadas pelo sistema foram:

a) **pronome:** possui pronomes interrogativos (qual, quais, que, quem, onde), possessivos (este, esta, esse, essa, estes, estas, isto, essas, esses, isso, meu, minha, seu, sua, aquele, aquela, aqueles, aquelas, aquilo) e indefinidos (muita, muitas, algum, alguns, alguma, algumas, vária, várias);

b) **artigo:** foram considerados os artigos mais utilizados pelo especialista do domínio de aplicação (o, a, os, as, um, uma, uns, umas);

c) **preposição:** possui as preposições, como por exemplo, ante, após, com, contra, em, entre, de, desde, para, sem, sob, sobre. Apresentando também junções de palavras, como dos (de + os) e do (de + o);

d) **numeral:** compõem todos os números na sua forma algébrica e escrita por extenso(1, 2, 3, 4, 5, 6, 7, 8, 9, 0, dois, três, quatro, cinco, seis, sete, oito, nove, dez);

e) **verbo:** os verbos estão descritos na sua forma completa, possuindo flexões de tempo e modo: é, tomar, estar, evoluir, foram, mostrar, apresentar, era, faz, ter,ser, submetidos, foi, estar;

f) **nomes:** esta classe é composta pelas denominações das doenças ou palavras referentes a área de urologia, dentre elas citam-se: paciente, pacientes, hospital, doença, evolução, remédio, dor vesical, dor em hipogástiro, febre, score prostático, toque, toque retal, nódulo próstata, cólica renal, cistite, biopsia, Hiperplasia Prostática Benigna, dentre outras expressões.

A composição dessas classes gramaticais considera as palavras utilizadas pelo médico especialista na sua rotina clínica. Na Figura 26 tem-se o pseudocódigo da análise léxica utilizado no Hades.

```
Lexico
Palavra[]:char
artigo[]:char
nome[]:char
pronome[]:char
numeral[]:char
inicio
enquanto palavra[] -> até comprimento do vetor , faça
se   ->  palavra[] = artigo então artigo
se -> palavra[] = pronome então pronome
se -> palavra[] = nome então nome
se -> palavra[] = numeral então numeral
se -> palavra[] = verbo então verbo
fim enquanto
```

Figura 26. Pseudocódigo Analisador Léxico

A Figura 27 representa os passos realizados durante a análise léxica na interface em processamento de linguagem natural para urologia – Hades.

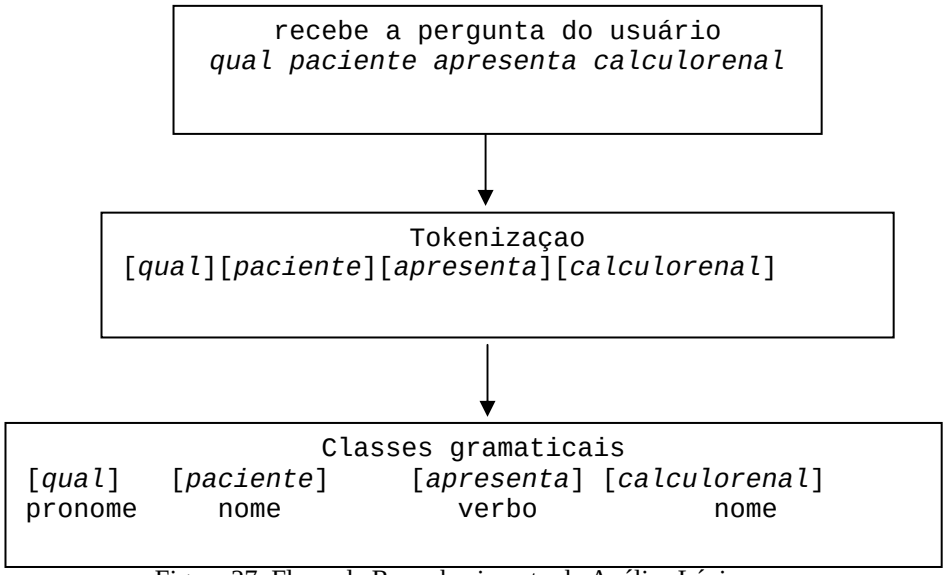

Figura 27. Fluxo de Reconhecimento da Análise Léxica

Finalizando a análise léxica, sendo todas as palavras aceitas pelo Hades, a próxima etapa (análise sintática) tem por objetivo, verificar a estrutura correta da sentença.

6.2.4 Implementação da Análise Sintática

A realização da análise sintática parte do uso de uma gramática que delimita a formação da frase segundo regras gramaticais da língua portuguesa. Assim, no Hades a formação das regras segue conceitos de uso para reconhecimento de linguagens por computadores (linguagens formais), mas adaptando para o uso de linguagem natural.

A gramática possui então, símbolos não-terminais e terminais, conforme descreve a Tabela 2.

Tabela 2. Símbolos Terminais e Não Terminais da Gramática

TERMINAIS	NÃO-TERMINAIS
verbo	FRASE <F>
pronome	SINTAGMA NOMINAL <SN>
preposição	SINTAGMA VEBAL <SV>
numeral	COMPLEMENTO NOME <CN>
nome	NUMERAL <NUM>
artigo	PREPOSICAO <PREP>
vazio (ε)	ARTIGO <ARTIGO>

Os símbolos terminais se referem a todas as classes gramaticais usadas pelo Hades, compreendendo nelas as palavras necessárias para o processamento. Os não-terminais referenciam as composições da frase na língua portuguesa.

O símbolo não-terminal <F> simboliza o início da frase, estando ela completa. Posteriormente subdivide-se em sintagma nominal e verbal. A partir de então, cada sintagma se particiona em um símbolo não terminal, gerando-se no caso do sintagma verbal um nominal como complemento ao verbo.

Na análise da oração usa-se a combinação dos terminais com os não terminais para o seu reconhecimento. Para isso, utilizou-se no Hades uma gramática livre de contexto (Figura 28), apresentando-se da seguinte forma:

```
<F> ::= <SN> <SV>
<F> ::= <SV>
<F> ::= <SN>
<SN>::= 'pronome' <F> <CN>
<SN>::= 'nome' <CN>
<SN>::= <ARTIGO>
<SN>::= <NUM>
<SN>::= <SV>
<SV>::= 'verbo' <PREP>
<SV>::= 'verbo' <SN>
<CN>::= <PREP> î
<CN>::= <SV>
<NUM>::= 'numeral'
<PREP>::= 'preposicao' <SN>
<ARTIGO>::= 'artigo' <SN>
```

Figura 28. Gramática do HADES

No algoritmo que implementa a análise sintática utiliza-se a gramática codificada, ou seja, atribui-se números aos símbolos terminais de 0 a 7 e para os não-terminais de 9 a 15. O número 8 não é utilizado, pois em linguagens formais o último elemento dos símbolos não terminais é empregado como fim de arquivo. Considerando-se que no desenvolvimento do Hades não usou um arquivo para fonte de dados de entrada, optou-se pela eliminação deste elemento. Na codificação da gramática foi utilizado o software GESPAS[19] que consiste em gerar tabelas e regras, ele foi desenvolvido pelo acadêmico Fabiano Martins.

Na realização da gramática livre de contexto seguiram-se alguns passos para obtenção das regras, ou seja, realizou-se a decomposição das orações em sintagmas. Considerando-se, por exemplo, a sentença: “Qual paciente possui HPB?”, tem-se a seguir os procedimentos realizados pela análise sintática no Hades.

Inicialmente a oração possui somente um estado chamado de Frase <F> (símbolo inicial da gramática), onde estão todas as classes e sintagmas. Posteriormente,

[19] Um software que se destina a geração de tabelas de parsing para analisadores sintáticos ascendentes e descendentes preditivos tabulares.

ocorre uma primeira subdivisão da oração, ou seja, a Frase <F>, passa a ter dois sintagmas um nominal e outro verbal (Figura 29).

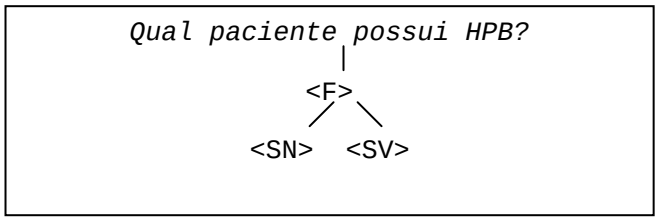

Figura 29. Inicio da Análise Sintática

Considerando-se o sintagma nominal pode haver a ramificação deste, em nós folhas, denominados de pronome e nome. Assim, finaliza-se o seu processo de análise (Figura 30).

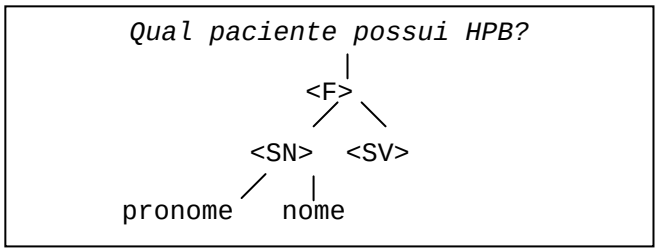

Figura 30. Reconhecimento do Pronome e Nome

Após a análise do sintagma nominal percorre-se a árvore até o outro sintagma existente, no caso o verbal, obtendo-se uma nova ramificação até um nó folha (verbo) e um símbolo não terminal <SN> (Figura 31).

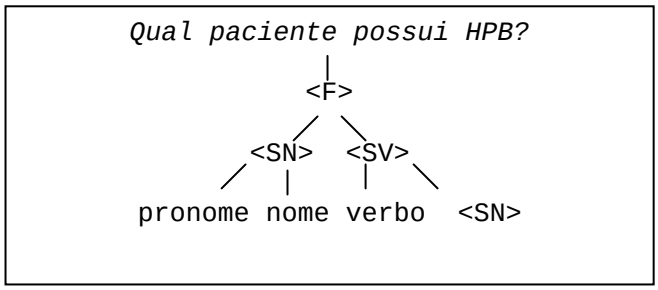

Figura 31. Reconhecimento do Verbo

Ao analisar o sintagma verbal, nota-se a existência de um novo símbolo não-terminal (SN) que deve ser derivado, possivelmente, em outro SN ou no caso da sentença analisada, em um terminal, chegando-se assim ao nó folha (nome) (Figura 32).

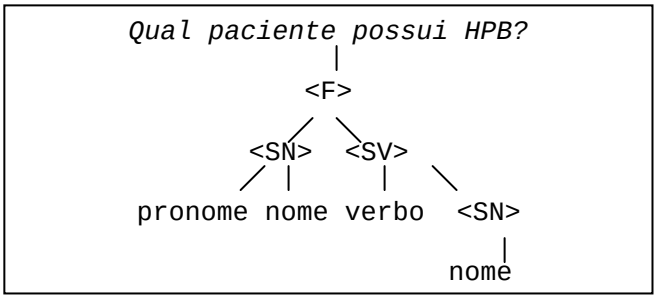

Figura 32. Reconhecimento do Nome

Na Figura 33, tem-se o pseudocódigo da análise sintática implementada no Hades.

```
Sintático
Variáveis
Palavra[] : char
Producoes[]: int
Parser[][] : int
 Equanto palavra [] -> até o comprimento do vetor, faça
         Se palavra[] = entrada léxico “palavra reconhecida”
         Senão
           int p = parser[palavra[][]palavras[][]]
           enquanto producoes [][] -> até o comprimento do
vetor
                  int h = produções[p]
             fim enquanto
         fim senão
fim enquanto
```

Figura 33. Pseudocódigo Analisador Sintático

Finalizada a análise sintática e sendo a sentença aceita pelo sistema, ocorre o encaminhamento dos resultados ao módulo de análise semântica e pragmática.

6.2.5 Implementação das Análises Semântica e Pragmática

Após a análise sintática e a verificação das regras gramaticais, tende-se a busca pelas palavras que dão o contexto geral da frase. Estas palavras referem-se às classes gramaticais (nomes, verbos e pronomes).

A partir da sentença, *Qual paciente possui HPB,* tem-se: a geração de uma árvore de derivação e posteriormente a realização da busca pelas palavras chaves a fim de encontrar: um agente (autor) relacionado ao nome *paciente,* este é ligado por meio do verbo *possui* a causa *HPB.* Esta análise gera um comando para encontrar a causa para um agente qualquer, na base de dados da área de urologia (Figura 35).

No desenvolvimento do Hades o módulo de análise semântica foi implementado juntamente com o pragmático, a fim de buscar a informação requerida pelo usuário diretamente da base de dados. Esta escolha deu-se em função da análise pragmática remeter ao contexto de mais de uma frase ou texto, porém como no Hades ocorre o reconhecimento de apenas uma frase, não se fez necessária a implementação da análise pragmática em um novo módulo, pois a semântica já possui o conhecimento sobre o contexto da frase. O pseudocódigo (Figura 34) da análise semântica e pragmática pode ser observado a seguir.

```
semânticopragmático
palavrachave[]
enquanto palavraschave[] -> até o comprimento da matriz, faça
             verificar as palavras chaves desta matriz
             palavrachave[]=nome
             palavrachave[]=verbo
             palavrachave[]=pronome
             fim verificar
         montagem do comando

palavrachave[nome](palavrachave[verbo]palavrachave[nome])
        fim montagem
fim enquanto
```

Figura 34 . Pseudocódigo Analisador Semântico-Pragmático

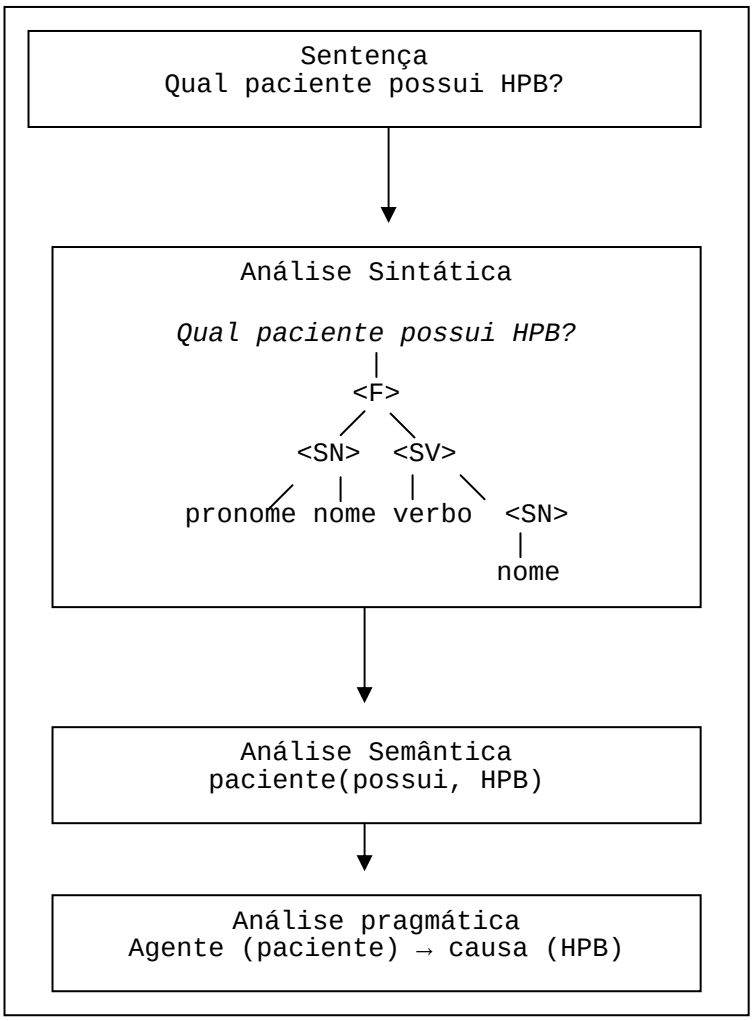

Figura 35. Fluxo de Análise Semântica e Pragmática

6.2.6 Implementação e Realização de Testes no Hades

A implementação do Hades foi realizada por meio da tecnologia Java, utilizando-se o ambiente de desenvolvimento Netbeans 5.5 que se encontra disponível para *download* em http://www.netbeans.org.

Na implementação do Hades, no que se refere a PLN, utilizou-se para embasamento teórico o livro de James Allen escrito em 1994 e intitulado de *Natural Language Understanding.*

A Figura 36 mostra a interface do Hades com menus e caixas de textos. Os menus possuem como função gravar a pergunta realizada pelo usuário e a resposta fornecida pelo sistema em um arquivo.txt. Nesta barra também se encontra a ajuda do sistema (Figura 37).

Figura 36. Interface do Hades

Figura 37. Ajuda do Hades

O menu de ajuda do sistema Hades foi desenvolvimento para os usuários conhecerem mais sobre a interface, e implementado em um ferramenta da Microsoft chamada de Microsoft Office Publisher, voltada para criação de páginas em html (Figura 38).

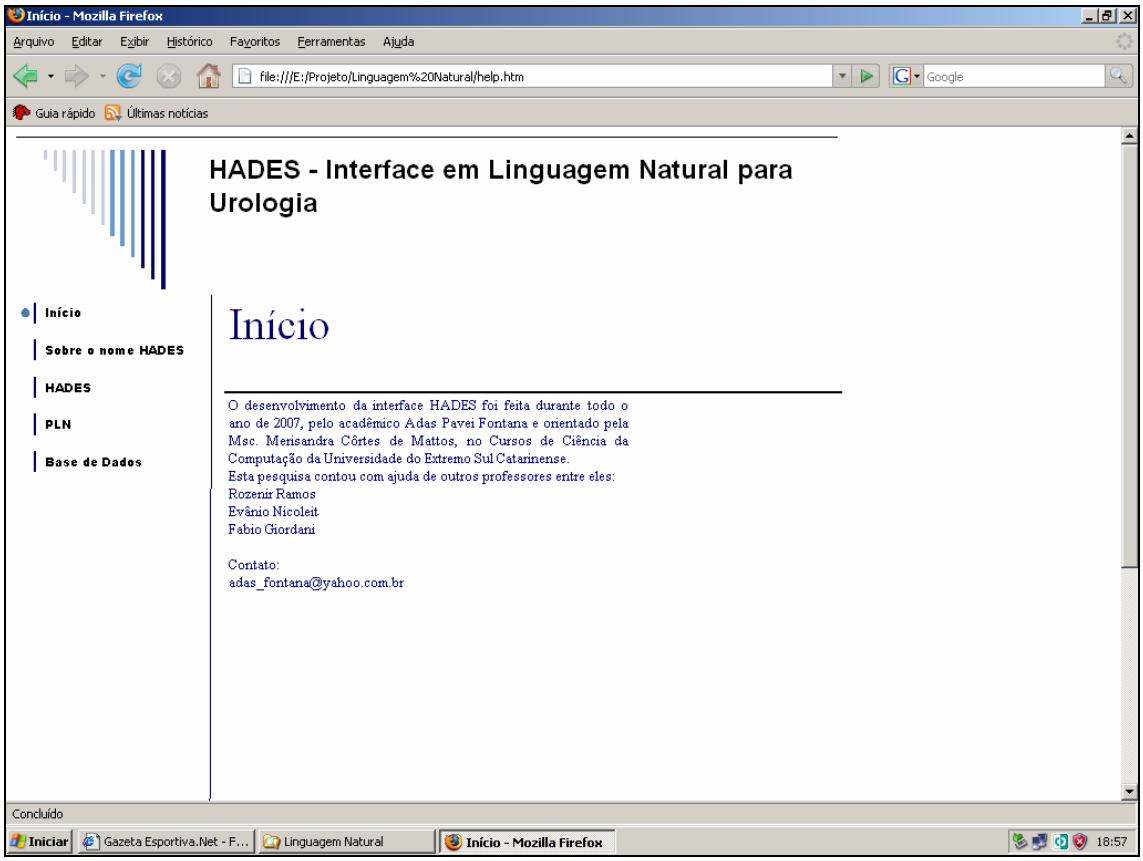

Figura 38. *Help* do Hades

Na formação dos dados de entrada, necessita-se que o usuário digite uma pergunta constituída por palavras presentes no dicionário. No processo de escrita de palavras compostas, como por exemplo, *câncer de próstata* o usuário deve escrever essa palavra sem o espaço em branco. Portanto, ela deve ser digitada como *câncerdepróstata* (devido ao processo de *tokenização* das palavras serem os espaços em branco). Ao concluir a digitação de seu questionamento o usuário não necessita digitar o ponto de interrogação no final da frase. No Hades as palavras podem ser escritas tanto em formato maiúsculo quanto minúsculo.

Na interface do sistema o botão que representa um ponto de interrogação referencia a pergunta e, quando acionado, inicia o processo de análise da frase. No campo de texto resposta retorna-se ao usuário o resultado da consulta, que no caso da

Figura 39, está trazendo apenas o código do paciente, pois informações de identificação desses não foram disponibilizadas devido às questões éticas e de privacidade.

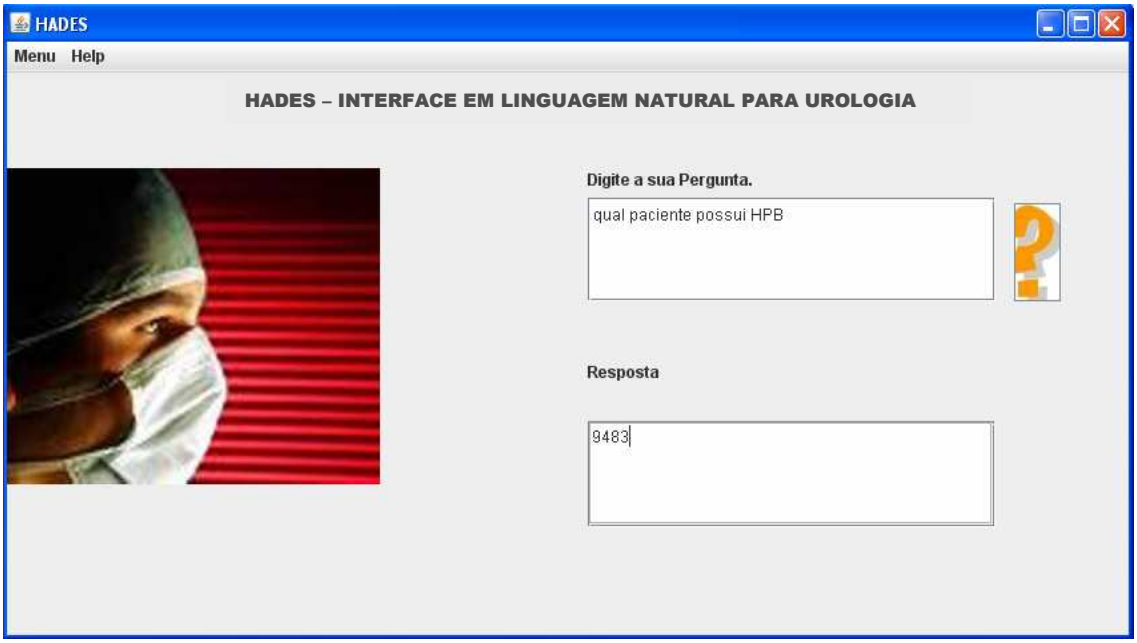

Figura 39. Entrada de Dados e Resultados.

Finalizando-se a implementação dos módulos de PLN no Hades, realizaram-se alguns testes por meio da base de dados da área de urologia.

6.3 RESULTADOS OBTIDOS

Os resultados foram obtidos pela realização de testes por meio de consultas na interface do Hades no que se refere a utilização da base de dados da área de urologia.

Na realização desses testes foram usados os seguintes requisitos de hardware: sistema operacional *Windows Professional Edition*, processador Intel 3.2 Ghz e memória *RAM* de 1 Gb.

A interface teve seu funcionamento testado por diferentes tipos de questionamentos referentes à condição do paciente, como por exemplo: *Quais pacientes apresentam cálculorenal?*, *Quais os sintomas apresentados pelo paciente que possui HPB?*, entre outros.

Na primeira sentença (Figura 40) o sistema se comportou de forma rápida e eficaz, conseqüência de possuir apenas o seguinte comando *paciente (apresentam, HPB)*. Na segunda sentença (Figura 41) com o acréscimo da condição de *sintoma*, que se agrega a doença *HPB*, o comando ficou um pouco mais complexo, apresentando-se da seguinte forma sintoma (paciente (possui HPB)).

Comparando-se estes dois questionamentos o processamento referente à primeira consulta torna-se mais rápido do que a segunda (aproximadamente 5s) pelo fato de não possuir uma outra agregação no sentido completo da frase.

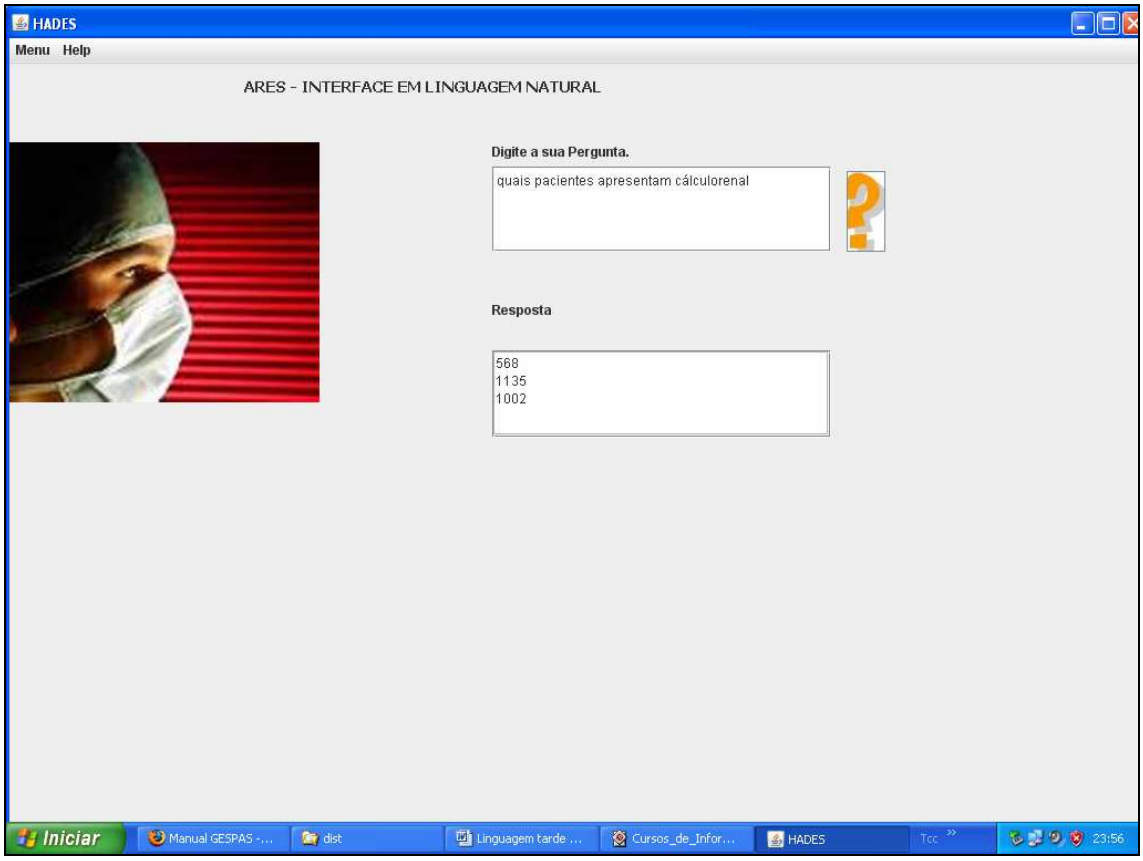

Figura 40. Saída no Hades para a primeira consulta

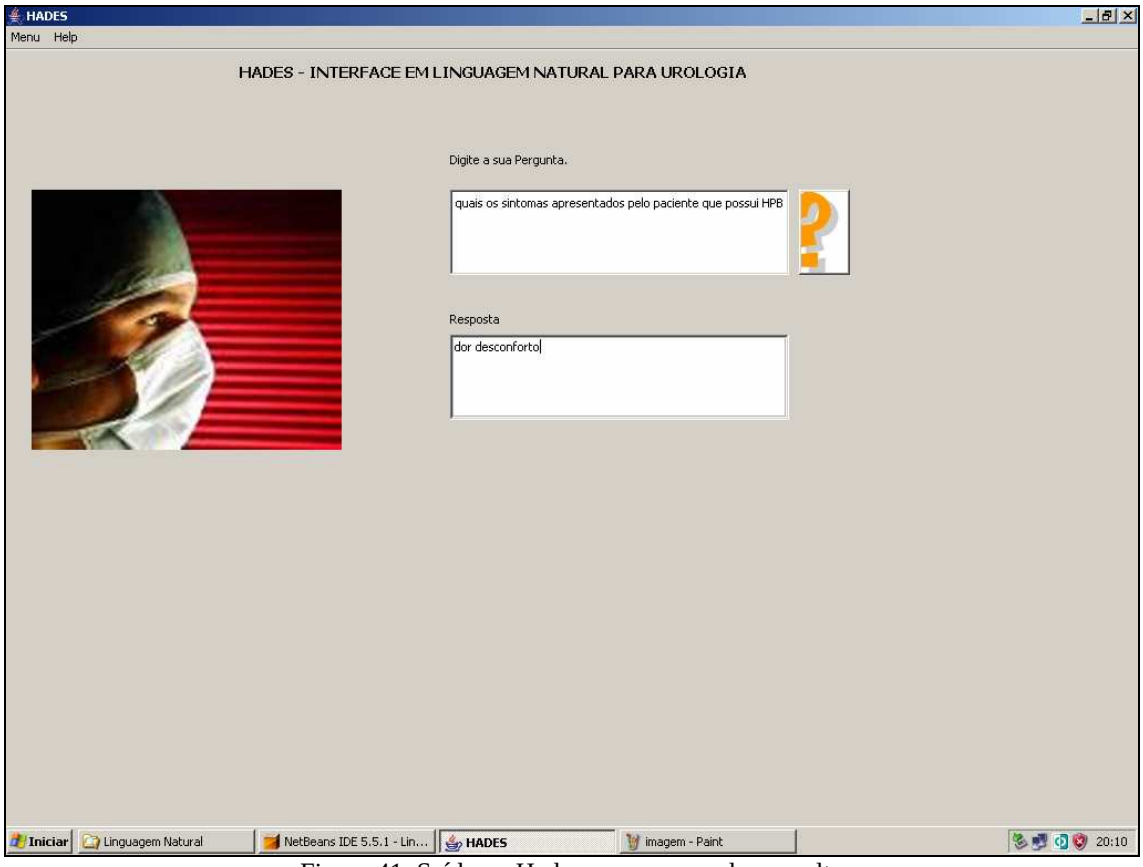

Figura 41. Saída no Hades para a segunda consulta

A partir dos testes, utilizando-se a base de dados bem como uma consulta com processamento de linguagem, verificou-se que a correta execução do algoritmo referente aos módulos de análises do Hades, repercutiu em um resultado claro e conciso.

CONCLUSÃO

As interfaces por meio de processamento de linguagem natural não são utilizadas apenas para melhorar aspectos visuais e usuais dos sistemas, mas também para facilitar a busca de informações em bases de dados, textos, entre outros. Assim, pode proporcionar benefícios no processo de interação homem-máquina.

Atualmente, com a crescente expansão da internet e da capacidade dos dispositivos de armazenamento de dados torna-se interessante o desenvolvimento de aplicações que integram processamento de linguagem natural e mineração de dados sendo que este possui como objetivo a descoberta de conhecimento a partir de grandes bases.

A aplicação do processamento de linguagem natural pode ocorrer por meio de extração, recuperação e tradução de textos, entre outros, utilizando-se diferentes métodos de análise. No caso desta pesquisa, realizou-se o desenvolvimento de uma interface que permite consultar informações de uma base de dados de maneira simples e intuitiva.

Na realização desta pesquisa verificou-se a importância e aplicabilidade dos conceitos de compiladores, no que se refere aos processos de análises léxica, sintática e semântica. Porém, constatou-se também a necessidade de bibliografias atualizadas e que abordem de forma mais detalhada o processamento de linguagem natural.

O Hades – Interface em Processamento de Linguagem Natural para Urologia obteve um desempenho satisfatório nas consultas realizadas, retornando respostas corretas, dentro do contexto da área de urologia e das informações disponíveis na base utilizada. Dessa forma, os objetivos da pesquisa voltados ao entendimento de PLN,

implementação das análises (léxica, sintática, semântica e pragmática), bem como ao desenvolvimento da interface em PLN foram atingidos.

Finalizando-se, têm-se algumas sugestões de trabalhos futuros a fim de dar continuidade a esta pesquisa e possibilitar novas abordagens para o processamento de linguagem natural, como por exemplo:

a) um enriquecimento da análise semântica e pragmática do Hades, bem como otimização do seu código;

b) aplicar essa interface em um Registro Eletrônico de Saúde (RES) que atenda todas as necessidades da base de Urologia;

c) desenvolvimento de *chatterbots* educacionais para o curso de ciência da computação;

d) elaboração de uma interface para banco de dados a fim de realizar consultas SQL;

e) utilização de algoritmos de aprendizagem para generalizar a informação extraída, gerando-se árvores de decisão por meio de algoritmos de aprendizagem como o ID3 e C4.5.

REFERÊNCIAS

ABRASCO. **Informações em saúde: acertos, erros e perspectivas**. In: CONGRESSO BRASILEIRO DE SAÚDE COLETIVA, 7., Brasília, 2003. Disponível em: http://www.abrasco.org.br/Congresso/20SC/Oficinas/Informac.pdf.

ALEXANDRINI, Fabio. **Desenvolvimento de uma Metodologia de Interpretação, Recuperação e Codificação Inteligente de Laudos Médicos Independentes de Idioma.** 2005. 100 f. Tese (Doutorado em Engenharia de Produção). Universidade Federal de Santa Catarina, Florianópolis.

ALLEN, J.F**. Natural Language Understanding.** Benjamin Cummings, 1987, Second Edition, 1994.

BARATA, Henrique Sarmento.; CARVALHAL, Gustavo Franco (Org.). **Urologia**: princípios e prática. São Paulo: Artmed, 1999.

BOURBEAU, L.; Carcagno, D.; Goldberg, E.; Kittredge, R.; Polguère, A. **Bilingual generation of Weather Forecasts in an Operations Environment**, 1990, Hans Karlgren (ed.), *Proceedings of COLING'90* (Helsinki, 1990).

CANTARELLI, Elisa M.P. **Acesso a Base de Dados através de Linguagem Natural.** 1998. 91f. Trabalho de Conclusão de Curso (Bacharel em Informática). Universidade Regional Integrada do Alto Uruguai e das Missões, Frederico Westphalen.

COELHO, Helder. **Inteligência Artificial em 25 Lições.** Porto: Orgal, 1995.

CONSELHO REGIONAL DE MEDICINA DO ESTADO DE SÃO PAULO. **Centro Biótico,** São Paulo, 1996. Disponível em: <http://www.bioetica.org.br/?siteAcao=Publicacoes&acao=detalhes_capitulos&cod_capitulo=13> Acessado em: 10 abril 2007.

COSTA, C. G. A. da. **Desenvolvimento e Avaliação Tecnológica de um Sistema de Prontuário Eletrônico do Paciente, Baseado nos Paradigmas da World Wide Web e da Engenharia de Software**. Dissertação (Mestrado em Computação) – UNICAMP, Campinas, 2001.

COSTA**,** Ernesto; SIMÕES, Anabela. **Inteligência artificial: fundamentos e aplicações.** Lisboa: FCA, 2004.

DATASUS. **Política Nacional de Informação e Informática em Saúde**. Disponível em: < http://w3.datasus.gov.br/datasus/datasus.php >. Acesso: 10 maio. 2007.

DIAS, Gonçalves. **Gonçalves** Dias **seleção de textos, notas, estudos biográfico, histórico e crítico e exercícios**. São Paulo: Abril Educação, 1982.

DIMARCO, Chrysanne; Foster, Mary Ellen. **The automated generation of Web documents that are tailored to the individual reader**, 1997 *Natural Language Processing for the World Wide Web, Papers from the 1997 AAAI Symposium* (Stanford, March 24-26, 1997), Menlo Park, California: AAAI Press

FERNANDES, Anita Maria da Rocha. **Inteligência artificial aplicada à saúde**. Florianópolis: Visual Books, 2004.

FIORIN, José L. **Introdução à Lingüística: Princípios de Análise.** 2 ed. São Paulo: Contexto, 2003.

FOWLER, Martin; SCOTT, Kendall. **UML essencial: um breve guia para a linguagem-padrão de modelagem de objetos.** 2.ed. Porto Alegre: Bookman, 2000.

HÜBNER, Jomi F. **Interface em Linguagem Natural**. 1992. 58 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação). Universidade Regional de Blumenau, Blumenau.

INCA – Instituto Nacional do Câncer. **Estimativa de Evidências de Câncer ano 2008.** São Paulo: 2007

IOM - Institute of Medicine. **The computer-based patient Record: an essential technology for health care**. Division of Health Care Services, Institute of Medicine, National Academy of Science, Washington, D.C., USA, 1997.

KRULEE, Gilbert K. **Computer Processing of Natural Language.** New Jersey: Prentice-Hall.

LANGACKER, Ronald W. **A linguagem e sua estrutura: alguns conceitos lingüísticos fundamentais**. 4. ed. Petrópolis: Vozes, 1980.

LUGER**,** George F. **Inteligência artificial: estruturas e estratégias para a resolução de problemas complexos**. 4. ed Porto Alegre: Bookman.

LOBATO**,** Monteiro. **Obras completas**. 16 ed. São Paulo: Ed. Brasiliense.

MENEZES, Paulo F. B. **Linguagens formais e autômatos**. 5.ed Porto Alegre: Sagra Luzzatto, 2005.

MINING, Cristopher D.; SCHUTZE, Hinrich. **Foundation of Statístical Natural Language Processing.** 2 ed. Massachusetts: Cambridge, 1999.

MICROSOFT. **MS Dewey.** Seatle: 2007.

MORAES, Ilara Hämmerli S. **Política, tecnologia e informação em saúde:** a utopia da emancipação. Salvador: Casa da Qualidade, 2002.

PESSOA**,** Fernando. **Alguma prosa**. Rio de Janeiro: Nova Fronteira, 1990.

POPOWICH, Fred. **Using Text Mining and Natural Language Processing for Health Care Claims Processing.** In. ACM SIGKDD Explorations Newsletter. Disponível em: <http://portal.acm.org/citation.cfm?id=1089815.1089824&coll=GUIDE&dl=GUIDE&CFID=15151515&CFTOKEN=6184618> Acessado em: 15 junho 2007.

PRESSMAN, Roger S. **Software engineering: a practitioner's approach**. 5.ed New York: McGraw-Hill, 2001.

PRICE, David; RILOFF, Joseph Z.; BRANDON, Harvery. **NaturalJava: A Natural Language Interface for Programming in Java.** In: Intelligent User Interfaces Conference, 2000, New Orleans. Anais eletronicos: New Orleans: ACM, 2000. Disponível em: http://www.informatik.uni-trier.de/~ley/db/conf/iui/index.html Acessado: 10 fev. 2007.

PRICE, Ana M. A. , TOSCANI, Simão S. **Implementação de linguagens de programação: compiladores**. Porto Alegre: Sagra Luzzatto, 2000.

RICH**,** Elaine. **Inteligência artificial**. São Paulo: McGraw-Hill, 1988.

RODRIGUES, Flávia B. V. **INSIRIUS – Interface Inteligente em Linguagem Natural para o sistema SIRIUS.** 1998. 90 f. Dissertação (Mestrado em Computação Aplicada). Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2004.

SACCONI, Luiz A. **Gramática Essencial da Língua Portuguesa.** 9 ed. São Paulo: Atual, 1992.

SACCER, N; LYMAN, M; BUCKNALL, C; NHAN, Tick LJ. **Natural language processing and the representation of clinical data.** In. National Center of Biotechnology Information Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=7719796&dopt=Abstract > Acessado em: 15 junho 2007.

SANTOS, Diana. **Toward language-dependent applications.** Machine Translation, 2001.

SBU – Sociedade Brasileira de Urologia. **Revista Oficial da SBU.** Botafogo: 2007.

SCHIPIURA, Cezar A. **CAS: uma interface em linguagem natural para auxílio à decisão na área da saúde.** 2004 192 f. Dissertação (Mestrado Tecnologia em Saúde). Pontifícia Universidade Católica do Paraná, Curitiba.

TAKEMURA, Tadamassa; ASHIDA, Nobuyuki. **A Study of the Medical Record Interface to Natural Language Processing.** In. ACM SIGKDD Explorations Newsletter. Disponível em: <http://portal.acm.org/citation.cfm?id=609079> Acessado em: 15 junho 2007.

TANIWAKI, C.Y.O. **Formalismos adaptativos na análise sintática de linguagem natural.** 2001.210f. Dissertação (Mestrado em Computação e Sistemas Digitais). Escola Politécnica da Universidade de São Paulo. São Paulo.

TERRA**,** Ernani. **Gramática & literatura: para o 2º grau.** 8. ed. São Paulo: Ed. Scipione, 1997.

TITONE, Renzo. **Psicolinguística aplicada introdução psicológica à didática das línguas**. São Paulo: Summus ed., 1983.

UFMT – Universidade Federal de Mato Grosso. **Sistema Integrado de Monitoramento Ambiental da Bacia do rio Cuiabá.** Cuiabá:2000

WANGENHEIM, Aldo V. Raiocínio baseado em casos. 1. ed Barueri, SP: Manole, 2003.

WIVES, Leandro Krug. **Um modelo de hiperdicionário: estudo de caso em prontuários médicos.** 1996. 68 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação). Universidade Católica de Pelotas, Pelotas.