
**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**
**CAMPUS CURITIBA**

**GERÊNCIA DE PESQUISA E PÓS-GRADUAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL - CPGEI**

**ROOSEWELT LEITE DE ANDRADE**

# DETECÇÃO DE ERROS EM TESAURO MÉDICO MULTILÍNGÜE ATRAVÉS DE CORPORA COMPARÁVEIS

**DISSERTAÇÃO DE MESTRADO**

**CURITIBA**
**DEZEMBRO DE 2006**

**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial

---

**DISSERTAÇÃO**
apresentada à UTFPR
para obtenção do Grau de

**MESTRE EM CIÊNCIAS**

por

**ROOSEWELT LEITE DE ANDRADE**

---

# DETECÇÃO DE ERROS EM TESAURO MÉDICO MULTILÍNGÜE ATRAVÉS DE CORPORA COMPARÁVEIS

---

Banca Examinadora:

Presidente e Orientador:

| | |
|---|---|
| **PROF. DR. PERCY NOHAMA** | **UTFPR** |

Co-orientador:

| | |
|---|---|
| **PROF. DR. STEFAN PAUL SCHULZ** | **UNI - FREIBURG** |

Examinadores:

| | |
|---|---|
| **PROF. DRA. CLÁUDIA M. C. MORO BARRA** | **PUC - PR** |
| **PROF. DRA. ANDREIA MALUCELLI** | **PUC - PR** |

Curitiba, dezembro de 2006.

**ROOSEWELT LEITE DE ANDRADE**

# DETECÇÃO DE ERROS EM TESAURO MÉDICO MULTILÍNGÜE ATRAVÉS DE CORPORA COMPARÁVEIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial do Centro Federal de Educação Tecnológica do Paraná, como requisito parcial para a obtenção do Grau de "Mestre em Ciências" – Área de Concentração: Engenharia Biomédica.

Orientador: Prof. Dr. Percy Nohama
Co-Orientador: Prof. Dr. Stefan Paul Schulz

Curitiba
2006

Ficha catalográfica elaborada pela Biblioteca da UTFPR – Campus Curitiba

A553d Andrade, Roosewelt Leite de
Detecção de erros em tesauro médico multilíngüe através de corpora comparáveis / Roosewelt Leite de Andrade. Curitiba. UTFPR, 2006
XIV, 91 p. : il. ; 30 cm

Orientador: Prof. Dr. Percy Nohama
Co-orientador: Prof. Dr. Stefan Paul Schulz
Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Curitiba, 2006
Bibliografia: p. 85 – 91

1. Engenharia biomédica. 2. Sistemas de informação. 3.Medicina – Tesauros. 4. Medicina – Vocabulário controlado. I. Nohama, Percy, orient. II. Schulz, Stefan Paul. Co-orient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. IV. Título.

CDD: 658.403811

# AGRADECIMENTOS

Aos meus orientadores, Profs. Drs. Percy Nohama e Stefan Paul Schulz, por todas as oportunidades, as orientações, paciência e aconselhamentos.

Agradecimento especial ao meu irmão e cunhada, Astrogildo Andrade Alves e Ana Rita Ioppi Alves. Meus queridos sobrinhos Isaac, Joana, Luana Alves e todos os outros da Família Ioppi, Daniela e "vó" Vanilda.

Aos companheiros de Freiburg, Kornel Markó, Philipp Daumke, Susanne Hanser, Olena Medelyan; Claudia Fink, Martin Schwarz, Oliver Osburg e o Jan Paetzold. Aos companheiros da área de saúde, Rafael Bruns, Viviane Seki Sassaki, Júnior Mendes, Maria Cláudia Hahn, Josiane Melchioretto, Thais Ariela Machado, Luciana Bandeira e Graziele Fátima Klein. Aos amigos da área de exatas, Jeferson Luis Bitencourt, Adriano Ricardo Duma, Michel Oleynik, Anderson Venturini, Ricardo Santos Guilherme Nogueira Neto, Píndaro Secco Cancian, Hood Wilson Gusso da Silva e o Prof. Edson José Pacheco. À Prof. PhD. Elisângela Ferretti Manffra, Viviana Raquel Zurro, as "meninas" da secretaria do PPGTS e ao amigo Márcio Luis Penkal e família pelos incentivos diários. Ao Sr. Jurair dos Santos, Sidnei Silva e William Dantas por me aturarem além do horário e todos os colegas do LER da PUCPR pelo convívio harmonioso.

Ao CNPq e ao DLR pelas bolsas e recursos disponibilizados. À Pontifícia Universidade Católica do Paraná pela infra-estrutura. Aos professores da Universidade Tecnológica Federal do Paraná.

Aos meus pais Raimundo Andrade da Silva (in memoriam) e Enedina Leite Andrade, pelos ensinamentos para a vida. Aos meus irmãos Stanley Leite de Andrade, Rooseleyde Leyde de Andrade e Rooseleny Leite de Andrade.

A Deus e à Santa Paulina.

“In the beginning there was information. The word came later.”
Fred I. Dretske

**SUMÁRIO**

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

ALUF - Albert-Ludwigs Universität Freiburg (Uni-Freiburg)
ANSI - *American National Standards Institute*
ASCII - *American Standart Code for Information Interchange*
ASK - *Anomalous State of Knowlegde* (Estado Anômalo do Conhecimento)
AVE - Acidente Vascular Encefálico
AvgP11 - Média dos Valores de Precisão sobre os 11 Pontos de Revocação
CID - Classificação Internacional de Doenças
CLIR - *Cross-Language Information Retrieval* (RI Multilíngüe)
CNPq - Conselho Nacional de Pesquisa e Desenvolvimento
COLING - *The Computatinal Linguistics Research Group Uni-Freiburg*
CUI - *Concept Unique Identifier*
DBMS - *Data Management System*
DLR - *Deutsche Forschungsanstalt für Luft-und Raumfahrt*
DOCS - Documentos
DTD - *Document Type Definition*
ECG - Eletrocardiógrafo
EUC - *Extended Unix Coding*
HTML - *Hyper Text Markup Language*
HTTP - *HyperText Transfer Protocol*
HTTPS - *HyperText Transfer Protocol Scheme*
IP - *Internet Protocol*
IRC - *Internet Relay Chat*
ISO - *International Standart Organization*
JDBC - *Java Database Connectivity*
JSP - *Java Server Pages*
LANG - Linguagem Indexada
LER - Laboratório de Engenharia de Reabilitação da PUC-PR
MB - Megabites

Medline - *MEDlars online*

MeSH - ***Me**dical **S**ubjects **H**eading*

MID - ***M**orphosaurus **ID**entifiers* (Descritores Semântico MS)

MS - ***M**orpho**S**aurus*

MSD - *Merck Sharp & Dohme Manual of Clinical Medicine*

MySQL - *My Structure Query Language*

NISO - *National Information Standards Organization*

NLM - *National Library of Medicine*

NLP - *Natural Language Processing* (PLN)

PLN - Processamento da Linguagem Natural

PUC-PR - Pontifícia Universidade Católica do Paraná

REQS - Expressão de Busca *(Query)*

RFC - *Requests for Comments*

RI - Recuperação de Informação

SGBD - Sistema Gerenciador de Banco de Dados

SGML - *Standart Generalized Markup Language*

SNOMED - *Systematized Nomenclature of medicine*

SRI - Sistema de Recuperação de Informação

SSH - *Secure Shell*

TCP/IP - *Transmission Control Protocol / Internet Protocol*

TREC - *Text Retrieval Conference*

UCS - *Universal Character Set*

UMLS - *Unified Medical Language System* (Sistema Unificado da Linguagem Médica)

URL - *Uniforme Resource Locator*

UTF - *Unicode Transformation Format*

UTFPR - Universidade Tecnológica Federal do Paraná

Web - Abreviação de *WWW*

WWW - *World Wide Web*

XML - *eXtensible Markup Language*

# RESUMO

A terminologia médica é complexa e esse fenômeno exerce um impacto forte na construção e manutenção de um tesauro do domínio médico. Metodologias para o controle de qualidade são de extrema importância, pois permitem detectar erros e consequentemente melhorar o desempenho de aplicações que utilizam tesauros, como, por exemplo, os Sistemas de Recuperação de Informações. Neste trabalho, propõe-se uma nova metodologia para a monitoração da construção e manutenção de um tesauro médico multilíngüe baseado em *subwords* através da utilização de *corpora* comparáveis para a detecção de descritores semânticos com problemas. Isso foi realizado comparando o perfil de distribuição de freqüência, em pares, dos descritores de um tesauro e verificaram-se os desequilíbrios na distribuição de ocorrências dos descritores semânticos para os idiomas português-inglês e alemão-inglês para serem corrigidos pelos lexicógrafos. Após as correções, uma avaliação sumativa foi realizada pela medida de parâmetro de desempenho que utiliza um *benchmark* de recuperação de informações padrão. A metodologia identificou problemas típicos como ausência de descritores semânticos, descritores diferentes com mesmo sentido, mesmo descritor com sentidos diferentes e ambigüidade dependente do idioma. Avaliando o desempenho na recuperação de informação, sobre o período do experimento, constatou-se um crescimento relativamente pequeno para os valores de precisão e revocação referente ao português e ao alemão. Houve um pequeno decremento para a língua inglesa, em contraste com o desempenho notável para a língua espanhola que alcançou um índice de 50%, em relação ao estado inicial dos valores de precisão, em três meses. Conclui-se que esse método é efetivo para a identificação de descritores com problemas e recomenda-se sua integração às operações de manutenção de um tesauro.

## *ABSTRACT*

Medical terminology is complex, a phenomenon which has a strong impact on the task of medical thesaurus construction and maintenance. A quality control methodology is therefore of utmost importance in order to detect errors in the thesaurus content, in order to improve the performance of applications using such a thesaurus, e.g. to support information retrieval systems. In this work, it is proposed a novel methodology to monitor the construction and maintenance of a medical multilingual subword thesaurus using comparable corpora to detect problematic semantic descriptors. By comparing the frequency distribution profile between thesaurus descriptors in pairs of comparable corpora, e.g. Portuguese-English and German-English, distribution imbalances were spotted and forwarded to the lexicographers which carry out the correction of the related thesaurus entries. After those corrections, a summative evaluation was done by measuring a performance parameter using a standard information retrieval benchmark. This methodology identified typical problems such as missing or dispensable descriptors, same sense in different descriptors, language dependent ambiguities. Evaluating the IR performance over time there was a relatively insignificant growth of the values for Portuguese and German. For English a minor performance decrease was detected. In contrast, the increment in performance of the Spanish part of the thesaurus was remarkable, since it amounted to a factor of more than 50% for three months. It is claimed that the proposed method is useful to identify weaknesses in a medical thesaurus and recommend to integrate it into the thesaurus maintenance workflow.

# CAPÍTULO 1

# INTRODUÇÃO

## 1.1 MOTIVAÇÕES

Desde há muito tempo, a humanidade produz, armazena e organiza as informações para serem recuperadas para quando houver necessidade (CARVALHO, 1999). De um modo geral, os dados e informações são representações de algum conceito que tem o objetivo de transmitir uma mensagem a um receptor. Da mesma forma, os seres humanos utilizam a cognição para a materialização mental das coisas do mundo real e utilizam símbolos para a transmissão de mensagens. No computador, a informação também pode ser, então, simbolizada em forma de texto, som, mídia ou imagem e símbolos podem ser utilizados para sua transmissão.

A proliferação de computadores por todo o mundo propiciou uma base sem precedentes para reunir a maior gama de símbolos entre as diversas culturas. De fato, isso gerou uma explosão de informações à disposição de qualquer pessoa ou máquina. Na era da Internet, a *Web* (*World Wide Web*) tornou-se a maior biblioteca do mundo.

Devido à intensa dinamicidade e à enorme quantidade de conhecimento em diversas áreas – e porque não se referir à banalização de informações, o uso de apoio computacional para a recuperação de informações textuais torna-se uma ferramenta obrigatória (HERSH, 1996).

Os prontuários eletrônicos, os artigos técnicos científicos, e outras publicações em mídia digital, da área de saúde, constituem-se numa vasta fonte de informações clínicas em formato textual. Porém, lidar com formato textual em processamento da linguagem natural não é fácil. O processamento de textos é complexo devido a sua diversidade de significados dependentes do contexto e outros fenômenos lingüísticos (FRIEDMAN e HRIPCSAK, 1999). De acordo com TAN (2001), mais de 80% das informações digitais encontram-se no formato textual; assim, torna-se importante que mecanismos de análise e processamento focalizem tal formato de informação, empregado nos documentos. Diante desse cenário, nas últimas décadas, vem ocorrendo progressos na área de Processamento da Linguagem Natural (PLN), pois assumindo que a informação seja primeiramente codificada como texto, a área de Recuperação de Informações (RI) é também um problema de PLN (STRZALKOWSKI,

1999). Essa área é considerada difícil por envolver outros conhecimentos devido a sua característica multidisciplinar.

Recuperação de informação é um processo de comunicação. Um Sistema de Recuperação de Informação (SRI) consiste de uma base de dados – onde são armazenados e disponibilizados dados, e um *software* para processar suas entradas e saídas. Na terminologia convencional de base de dados, os itens da base de dados são chamados de registros. Na terminologia de RI, entretanto, os registros são chamados de documentos e, portanto, pode-se chamar de base de documentos para a área de RI (HERSH, 1996).

A recuperação é um processo de interação com o SRI com o objetivo de obter documentos considerados relevantes ou não, num certo domínio, para uma determinada necessidade de informação. O termo recuperação de informação é um termo amplo e ainda não plenamente definido, assim como o próprio termo "informação". Apesar do usuário interagir com o sistema devido a uma necessidade de informação, o inverso não acontece, ou seja, um SRI não informa o usuário sobre o assunto relacionado ao seu questionamento – não há mudança de estado de conhecimento. Um SRI somente informa sobre a existência ou não dos documentos relacionados à sua requisição. E, desta forma, a qualidade de um sistema de recuperação de informação depende tanto da proporção de documentos recuperados dentre o total considerado relevante, nomeadamente revocação *(recall)*, quanto do grau de exclusão de documentos irrelevantes, chamada de precisão *(precision)*.

Em RI existe uma razão prática para considerar aspectos filológicos[1] ou filosóficos das palavras. E o cerne desta questão pode ser verificado no seguinte axioma de Meadow (MEADOW, BOYCE e KRAFT, 1992):

> *"For any given message or text, the determination of whether it is a data or information, or contains news or wisdom, is in the mind of beholder and not in the recorded symbols".*
>
> *"Para uma dada mensagem ou texto, a determinação do que é um dado ou informação, se contém notícia ou transmite sabedoria está em poder do observador e não somente nos símbolos da mensagem ou texto".*

[1] Filologia refere-se a um conjunto de conhecimentos necessários para interpretar e conhecer um texto, que antigamente se ocupava em fixar e comentar os textos literários, procurando extrair regras de uso lingüístico e que, modernamente, estuda a língua, a literatura e todos os fenômenos culturais de um povo por meio dos seus textos escritos; distinguindo-se, no entanto, da lingüística, na medida em que esta centra o seu interesse na língua, e aquela nos textos.

De fato, na prática, quem realmente detém o poder e a capacidade de julgamento do que é dado e o que é informação é o próprio ser humano.

A área de RI textual pode ser classificada como RI monolíngüe ou RI multilíngüe (*Cross-language Information Retrieval - CLIR*) (OARD, 1997; PETERS, 2000). A diferença entre um SRI multilíngüe e um SRI monolíngüe é a habilidade do sistema multilíngüe recuperar documentos em uma língua natural diferente da utilizada na consulta.

Existem basicamente dois processos envolvidos na RI: indexação e recuperação que, por sua vez, podem ou não estar suportadas por um tesauro.

O tesauro é um conjunto de termos relacionados entre si, com sinônimos e relações semânticas, utilizadas para representar conteúdos de documentos, com a finalidade de classificação ou busca de informação (CINTRA, 2002). A idéia principal de se utilizar um tesauro é prover um vocabulário controlado de referência a um SRI (FOSKETT, 1997). Com o auxílio de um tesauro, pode-se indexar e recuperar documentos em um determinado domínio.

A construção de um tesauro envolve alguns passos. Basicamente, o primeiro é definir o domínio de atuação. Uma vez definido e delimitado tal domínio, o passo seguinte será compilar um *corpus*[2] de termos representativos da terminologia do domínio, de tal forma que seja a matéria prima para a construção do tesauro proposto (SOERGEL, 1997). Não há critérios objetivos para determinar a representatividade. Quando se diz que um *corpus* deve ser representativo, entende-se representatividade em termos de extensão do *corpus*, isto é, de uma quantidade determinada de palavras e de textos. A nomenclatura empregada na Lingüística de *Corpus* para definir o conteúdo e o propósito dos *corpora* é extensa. Os principais tipos de *corpus* citados na literatura são agrupado segundo critérios de: modo, tempo, seleção, conteúdo, autoria, disposição interna e finalidade. Os *corpora* montados nesse trabalho classificam-se como de seleção e amostragem (*sample corpus*) e de finalidade estatística: construído para permitir o desenvolvimento de aplicações e ferramentas de análise. Neste trabalho, os *corpora* foram utilizados para o processo de alinhamento de forma a detectar discrepâncias de ocorrências de MIDs, que representam coisas do mundo real ou abstrato, entre idiomas. Alinhar é realizar verificações explícitas de correspondências entre segmentos (semânticos e/ou sintáticos) de uma língua em relação à outra. O alinhamento não depende obrigatoriamente de um processo de etiquetagem das palavras, mas uma segmentação prévia é sempre necessária (SARDINHA, 2004).

[2] Em lingüística, *corpus* é uma coleção de textos. *Corpora* é uma coleção de *corpus* – e nesse trabalho, cada *corpus* é uma coleção de textos, compilados do domínio médico, de um idioma distinto.

A importância de usar um tesauro decorre do fato que grande parte da informação é criada e expressa por meio da linguagem natural. Isso acontece porque a linguagem natural representa o modo de comunicação dos seres humanos, onde se utilizam diferentes vocabulários para expressar suas intenções (FURNAS, 1987) através de mensagens - elemento material através do qual um conjunto de informações, organizadas segundo um código, circula entre um emissor e um receptor. A diversidade da linguagem humana (a mesma idéia pode ser expressa por múltiplas expressões lingüísticas) dificulta o uso de técnicas de RI. Além disso, como objeto de inferência humana na sua construção, sujeito a erros. Desta forma, é necessário que haja meios de avaliar a representatividade do tesauro diante do sistema, pois este também estará utilizando algum tipo de abordagem implementada nas suas heurísticas de processamento. A avaliação poderá ser realizada de maneira formativa ou sumativa. Avaliação formativa é um método de julgamento realizado durante a evolução do processo – enquanto as atividades estão ocorrendo; é focado no processo. A avaliação sumativa é o método de julgamento realizado ao final dos processos. O enfoque encontra-se nos resultados finais (BHOLA, 1990).

Diferentes componentes são associados com o entendimento de uma mensagem. Os mais comuns são a sintática, a semântica e o domínio do contexto. Nessa tríade, pode-se situar a área de PLN como um intermediador a ser utilizado para o entendimento com base em um modelo conceitual de um domínio ou um vocabulário controlado da terminologia médica. Assim, o uso de um vocabulário controlado, tal como fornecido por um tesauro, pode melhorar o resultado de RI em larga escala, já que um vocabulário controlado de terminologia médica melhora a RI de documentos médicos, pois cada conceito do vocabulário está estritamente associado ao seu significado de fato e às suas acepções restritas ao domínio; reduzindo, assim, a variedade e a ambigüidade (FRIEDMAN e HRIPCSAK, 1999).

Termos são vocábulos relacionados aos seus conceitos, previamente definidos, peculiar a um domínio. Assim, de forma textual, as palavras são unidades mínimas com som e significado que, por si só, podem constituir enunciado, forma livre ou lexema.

São diversos profissionais da área de saúde que utilizam jargões e outros termos específicos de cada especialidade. Devido à riqueza de expressões – provavelmente mais do que em outros domínios - a implementação de sistemas que lidam com linguagens naturais torna-se complexa quando o objetivo é realizar buscas orientadas a conceitos ou sentidos. Diante desse cenário, um sistema de recuperação de informações precisa ter suporte em um tesauro, ou seja, um vocabulário controlado que responda a essas questões (SCHULZ e HAHN, 2000).

Os SRI normalmente baseiam-se em tesauros e devido a fenômenos lingüísticos, o processo de criação e manutenção torna-se complexo. As formas gráficas que constituem as palavras de um texto (*tokens*) são muitas vezes ambíguas, podendo freqüentemente uma mesma forma corresponder a diferentes flexões de duas ou mais entradas lexicais distintas. Esse fato, aliado a uma abordagem por uma representação artificial de um dado conhecimento torna o trabalho mais interessante e complexo. E neste trabalho, utilizaram-se descritores atomicamente semânticos mapeados para uma representação independente do idioma chamada de MID (***M****orphosaurus* ***ID****entifier*). Equacionar questões relacionadas à delimitação sintática dos termos, relevância lexical, relevância semântica ou, até mesmo, relevância conceitual, na implementação de um tesauro, não é uma tarefa fácil.

Esses problemas devem-se ao fato de que os termos não são simplesmente palavras, mas uma unidade com carga semântica inserida num contexto específico, realizado por seres humanos. E, como tal, sujeito a erros. Esses erros geram ruídos num SRI com um todo e precisam ser tratados.

Desta forma, conclui-se que é necessário examinar esses problemas à luz da abordagem adotada, corrigindo-os de forma a produzir um adequado desempenho num SRI específico de um domínio. Assim, para assegurar a recuperação de um número desejável de documentos relevantes e garantir uma seleção mais precisa, deve-se fazer um controle da terminologia, que delimite os meios pelos quais poder-se-á expressar idéias, não necessariamente estabelecer limites, mas sim, regras que permitam a expansão e efetividade do sistema através de bom controle vocabular, que garanta efetividade nas relações entre perguntas e respostas (JESUS, 2002).

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

A qualidade do tesauro e, conseqüentemente, a diminuição do ruído (perturbação) num SRI é função do equacionamento de questões relacionadas aos fenômenos lingüísticos e aos problemas causados por heurísticas de implementação do próprio SRI. Como se trata da construção de um tesauro com a inferência humana, é de se esperar erros que podem ou não ser sistemáticos.

Desta forma, propõe-se neste projeto de pesquisa uma metodologia cujo objetivo é implementar uma sistemática para monitorar a criação e a manutenção de um tesauro por

meio da comparação de ocorrências de descritores semânticos bilíngües gerados a partir de *corpora* comparáveis, nas línguas portuguesa, alemã, inglesa, espanhola e sueca, com vistas ao incremento do desempenho num SRI em saúde.

Assume-se que a discrepância entre as ocorrências de descritores semânticos normalizados entre línguas seja um indício de potencial problema num sistema de vocabulário controlado multilíngüe.

### 1.2.2 Objetivos específicos

A pesquisa pode ser dividida em duas grandes etapas: (a) geração de *corpora* multilíngüe no domínio da saúde, visando a construção de listas de ocorrências de descritores semânticos normalizados para análise e, se necessário, a correção dos descritores semânticos bilíngües normalizados e, finalmente, (b) avaliação do desempenho da metodologia.

Para alcançar o objetivo geral descrito, é necessário realizar as seguintes tarefas:

*(primeira parte)*

(1) montar *corpora* nas línguas inglesa, portuguesa, alemã, espanhola e sueca;
(2) mapear o conteúdo textual de cada *corpus* para descritores semânticos;
(3) organizar, em ordem decrescente, as freqüências de ocorrências de MIDs bilíngues;
(4) analisar as primeiras 160 MIDs seguindo a ordem de classificação;
(5) realizar *backups* diários do tesauro para montagem posterior das curvas de precisão e revocação;

*(segunda parte)*

(6) preparar *workbench*[3] para o processamento das curvas de precisão e revocação;
(7) plotar as curvas de precisão e revocação para cada uma das dez versões de tesauro no período de correções das MIDs;
(8) analisar os resultados.

## 1.3 ESTRUTURA DA DISSERTAÇÃO

Esse documento está estruturado da seguinte maneira: no capítulo 2, apresenta-se o estado da arte referente à Recuperação de Informações. Nos seus sub-capítulos, descrevem-se conceitos relacionados à representação do conhecimento, a área de Recuperação de

[3] *Workbench* é definido como um ambiente que contém um conjunto de ferramentas computacionais para a automatização de um processo completo para a geração de um resultado.

Informações, seus modelos e forma de avaliação. Depois, explana-se sobre vocabulário controlado seguido de um tópico sobre Lingüística de *Corpus*. Então, detalhadamente, apresentam-se as especificações do tesauro do sistema *Morphosaurus* utilizado como *Workbench*.

A montagem do *workbench* para a realização dos procedimentos é uma tarefa que exige conhecimentos da área de computação para a implementação de ferramentas necessárias a processamentos lingüísticos. Além da infra-estrutura necessária, o capítulo 3 trata da metodologia empregada neste trabalho, ou seja, a forma como foram gerados os *corpora* estatísticos as listas de ocorrências bilíngües de MIDs, os procedimentos executados pelos lexicógrafos e, finalmente, os passos necessários para avaliação da metodologia pela evolução das médias dos valores de precisão sobre os onzes pontos de revocação (AvgP11) .

No capítulo 4, referente aos resultados gerados, são apresentados os tipos de problemas encontrados no tesauro, que só é possível categorizá-los após a análise dos mesmos. Finalmente, aborda-se o desempenho obtido pela aplicação da metodologia criada e a evolução das médias AvgP11 do tesauro multilíngüe.

No Capítulo 5, são discutidos os resultados encontrados na análise dos resultados numéricos. Também são apontados motivos para justificá-los. Finalmente, descrevem-se as principais contribuições trazidas pela pesquisa realizada e seus futuros desdobramentos, os quais poderão complementá-la, aprofundá-la e expandir o presente estudo.

# CAPÍTULO 2

# FUNDAMENTAÇÃO TEÓRICA

## 2.1 INTRODUÇÃO

A *Web* propiciou tanto a explosão quanto a banalização e a globalização das informações, o que levou ao desenvolvimento de uma área impar chamada Recuperação de Informações Multilíngüe – Cross Language Information Retrieval (CLIR), e que pode ser vista como uma intersecção entre a área de RI e a lingüística relacionada à tradução – máquinas tradutoras (*Information Retrieval and Machine Translation*), onde ambas compartilham de problemas específicos. Esta nasceu da necessidade de traduzir o texto para uma outra língua e recentemente recuperar documentos em outras línguas que fazem parte do mesmo contexto. Elas nasceram bem antes de existir a *Web* (GREFENSTETTE, 1998).

O "Problema da Recuperação de Informação" que vem sendo estudada há inúmeros anos pode ser descrita como:

> *"um modo na qual pode-se distinguir uma informação relevante de uma informação irrelevante para satisfazer a uma certa necessidade de informação(RIJSBERGEN, LALMAS e HUIBERS, 1996)".*

Existem vários modelos de SRI, entre eles pode-se citar os clássicos modelos Booleano e Espaço Vetorial (SALTON, 1971), além do modelo probabilístico introduzido por S. E. Robertson e Spark Jones, em 1976. Mais recentemente, em 1986 Rijsbergen (RIJSBERGEN, LALMAS e HUIBERS, 1996) propôs um modelo de RI baseada na lógica. Nesta proposta, defende-se que a lógica é uma base que pode prover uma escala de conceitos poderosa muito útil para a modelagem de documentos e expressão de busca para os propósitos da RI.

Uma variedade de abordagens tem sido utilizada em RI variando em escopo e domínio. A delimitação em um determinado domínio é importante para que se possa, em primeiro lugar, diminuir problemas gerados por ambigüidades advindas de interpretações sintáticas ou semânticas inerentes ao processo da linguagem natural e, conseqüentemente, em segundo lugar, melhorar a performance de um motor de busca (HERSH, 1996).

Um sistema é uma combinação de componentes que atuam conjuntamente e realizam um certo objetivo. O conceito pode ser aplicado inclusive a fenômenos abstratos. Um "ruído" ou uma perturbação (ou distúrbio) é um sinal que tende a afetar adversamente o valor da saída do sistema, do resultado final. Um sistema mantém uma relação prescrita entre saída e alguma entrada de referência comparando-as e utilizando a diferença como um meio de controle; sendo denominado sistema de controle realimentado (OGATA, 1990). Um SRI pode ser visto como composto por vários componentes, entre os quais o motor de busca, com sua heurística para a ordenação dos documentos selecionados, um vocabulário controlado, um módulo para processamento da linguagem natural, etc... Cada um desses componentes contribui com uma cota de "ruído" no sistema. A diminuição do ruído no sistema está intrinsecamente ligada à boa construção, implementação, configuração, etc..., desses componentes, enquanto pertencente à engrenagem. Isso se traduz em qualidade dos componentes envolvidos. Um componente que não produz ruído num dado sistema pode ser causa de mau desempenho em outro. Num tesauro, classes de equivalências mal definidas, relacionamentos semânticos mal configurados, considerações sobre relevâncias lexicais podem ser comparados como sinais que podem provocar perturbações num SRI como um todo. A qualidade aqui tratada refere-se basicamente a sua boa representatividade dos diversos significados da terminologia de um determinado domínio, não levando em conta aspectos técnicos de construção e nem sua estruturação definidas em normas[4].

## 2.2 CONCEITOS

Este documento utiliza-se de diversos conceitos que são interpretados, algumas vezes, de maneira incorreta ou, por vezes, de um outro modo devido ao fato de possuir outros significados, dependendo da área ou do contexto nos quais se inserem. Para evitar tais problemas, devido à sua natureza ambígua, optou-se por descrever, mesmo que de maneira superficial, como cada um dos conceitos envolvidos neste trabalho é considerado.

### 2.2.1 Dado

**Dado** é uma *string* de símbolos elementares. Não precisa existir um significado para

[4] A ANSI/NISO Z39.19-2005 é um norma para a construção, formatação e gerenciamento de vocabulários controlados monolíngüe.

todos os símbolos, mas precisa estar claro que o atributo do dado é um valor (MEADOW, BOYCE e KRAFT, 1992). O dado consiste de um resultado da observação e é uma medida acerca das coisas do mundo real.

Formalmente, um dado constitui-se de uma representação simbólica de um objeto ou informação pertencente a um domínio, sem levar em conta considerações de contexto, significado ou aplicação (ABEL, 2001).

Muitas pessoas sabem que existe diferença entre dado e informação, mas normalmente esses termos são utilizados como sinônimos, pois eles não sentem a necessidade de fazer distinção numa conversa do dia-a-dia (MEADOW, BOYCE e KRAFT, 1992). Embora ninguém tenha arriscado igualar os dois conceitos, por questões de praticidade, neste trabalho, "dado" é informação.

### 2.2.2 Documentos

Neste trabalho, o termo *documento* é utilizado para denotar um registro textual, em linguagem natural. Em um estudo realizado por (Michael Buckland, 1997), da Universidade da Califórnia, foram coletadas as seguintes definições para documento:

a) "qualquer base material capaz de estender nosso conhecimento, que seja disponível para estudo ou comparação, pode ser um documento" (WIVES, 2004);
b) "um documento é uma evidência que suporta um fato. [...] qualquer signo físico ou simbólico, preservado ou registrado, com a intuição de representar, reconstruir ou demonstrar um fenômeno físico ou conceitual é um documento" (WIVES, 2004).

### 2.2.3 A Informação

Para saber o que é RI, primeiramente, deve-se saber o que é informação. A rigor, não existe uma definição satisfatória. A noção de informação é vista de várias maneiras por várias pessoas. O dicionário Webster (GOVE, 1986) possui sete definições sobre a informação, entre as quais, citam-se: "comunicação ou recepção do conhecimento ou inteligência", "fatos ou figuras utilizados na comunicação que são distintamente organizados formalmente para representar um conhecimento", ou, "a forma como um objeto do conhecimento é formado na mente para transmitir um estado ou evento do mundo real" ou ainda, "uma medida quantitativa da incerteza do resultado de um experimento". Por enquanto, simplificam-se as

características da informação como sendo algo que (a) é representado por um conjunto de símbolos que (b) são organizados dentro de uma estrutura, e (c) que podem ser lidos e entendidos.

A informação deve ser entendida como um "conteúdo", separado de qualquer suporte físico, livro, vídeos, etc., pois segundo (MIRANDA, 1996), a informação independe de seu suporte, isto é, ela não depende de registro material para existir e, por este motivo, requer novas abordagens teóricas e metodológicas, novas práticas e novas tecnologias para seu ciclo de vida e transformação.

Vários modelos matemáticos foram desenvolvidos para expressar a geração, transmissão e a utilização da informação. Muito dos aspectos teóricos sobre a informação podem ser encontrados nos trabalhos realizados por (LOSEE, 1990).

Muitos cientistas creditam a teoria da informação aos trabalhos de Claude Shannon e Warren Weaver (SHANNON e WEAVER, 1949). Suas maiores contribuições foram a técnica de codificação e a decodificação de sinais, assim como a minimização do ruído introduzido no sistema – figura 1. Weaver, por outro lado, concentrou- se em estudar o significado da informação e de como esta poderia ser transmitida.

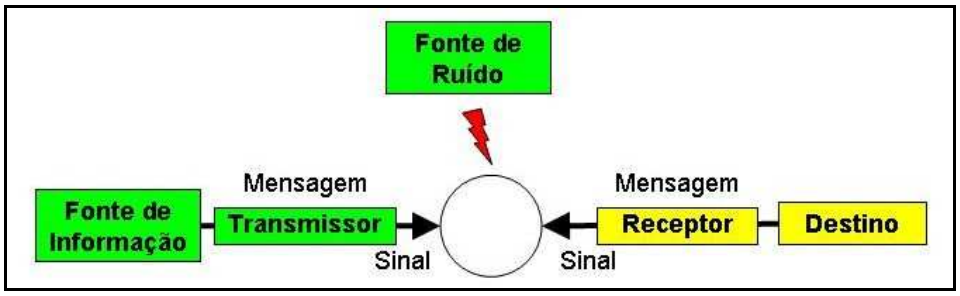



Figura 1: Diagrama do Modelo de Comunicação de Shannon e Weaver.

Do ponto de vista do transmissor, o objetivo é enviar a informação de modo eficiente e mais compreensivo possível. Entretanto, **a informação é uma medida da incerteza ou medida da entropia**. Shannon definiu quantitativamente essa medida através de uma fórmula muito simples, expressa na equação (1).

$$Informação = H = \log(1/p) = -\log(p) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

onde:

- $p$ é a probabilidade da ocorrência de um símbolo numa mensagem no sistema;

- $N$ é a quantidade de símbolos utilizados por um idioma ou o sistema de codificação utilizado no sistema.

No sistema alfabético, se cada letra possui a mesma probabilidade de ocorrência, então a chance de qualquer letra ocorrer é de 1/26. A informação contida em cada letra é $-\log(1/26) = 4{,}7 bits$. Por outro, a informação gerada por cada rodada num jogo de moeda "cara-coroa" (*coin flip*) mede $-\log(1/2) = 1 bit$. Conclui-se que existe mais informação numa letra que num numa lance do jogo "cara-coroa".

Em uma dada linguagem natural, cada caractere possui uma probabilidade associada de ocorrência e, normalmente, não se repetem. Se dois idiomas utilizarem o mesmo alfabeto, como por exemplo, inglês e francês, pode haver termos com freqüências diferentes para o mesmo texto – e é o que ocorre! Essa medida poderia servir de apoio à decisão para escolher a mensagem ou sistema de transmissão a ser utilizado, por exemplo. *H* não é uma medida de conteúdo da informação de um texto isolado, ou uma palavra ou uma mensagem. Mas o trabalho de Shannon representou o começo de uma ciência formal para uma medida da informação.

Shannon estava interessado em medir a quantidade de informações que podia ser enviada por um canal de comunicação. Utilizando uma linguagem em que todas as palavras possuem igual probabilidade de ocorrência, a taxa de informações enviadas por palavras é menor que uma outra linguagem que possui uma faixa maior de probabilidade de ocorrências das palavras (caso da linguagem natural). Por exemplo, o artigo "*o*" não traz muita informação; desta forma, perde-se muito na sua transmissão. Essa é a razão do porque os já ultrapassados sistemas de telegrafia, ou estilo de cabeçalhos de jornais (*newspaper headline style*), não utilizavam artigos, ou palavras muito comuns, nos cabeçalhos; entretanto, textos completos precisam desses artigos, tanto para a precisão do significado quanto do estilo. Em vocabulários controlados, a mesma estratégia é utilizada, por exemplo, dando peso a algumas classes de lexemas. Um tipo de problema encontrado durante os experimentos trata-se da delimitação do tamanho de uma *string*, onde em alguns casos foi necessário acrescentar ou retirar uma letra para manter a boa segmentação e assim manter o significado correto do termo artificial gerado.

Em uma definição operacional, *a informação é um dado que pode mudar o estado de percepção de um sistema, seja de um computador, seja de um cérebro* (MEADOW, BOYCE e KRAFT, 1992).

Uma outra definição relacionada expressa que *a informação é aquilo que é utilizado para inferir numa decisão*. A informação como valor para tomada de decisão está amarrada ao conceito de redução da incerteza. Informação é termo polissêmico.

A fórmula de Shannon é uma ferramenta válida de medida da informação em resposta aos problemas de engenharia encontrados, relacionados à transmissão de mensagens via meios eletrônicos. Mas outra questão surgiu: se essa medida poderia ser aplicada em outras áreas, especialmente na área de transmissão de informações médicas. Herckerling (HECKERLING, 1990) utilizou a teoria de Shannon para demonstrar que a informação, utilizando testes com diagnósticos de prontuários médicos, baseados na probabilidade de ocorrências de doenças, freqüentemente, era insuficiente para apresentar diagnósticos de prontuários semelhantes.

Outros trabalhos foram realizados no sentido de melhorar o modelo de Shannon e Weaver. Bar-Hillel e Carnap (BAR-HILLEL e CARNAP, 1953) incrementaram uma camada semântica à medida da informação. Descobriu-se que a informação não trata somente de ser uma seqüência isolada de bits, mas objetos ligados por relacionamentos. Esses objetos e relacionamentos podem ser codificados de forma lógica, de modo a definir a informação como um conjunto de manifestações, tornando-a mais precisa.

### 2.2.4 Conhecimento

De um modo geral, conhecimento parece representar um alto grau de certeza, de convicção, do que propriamente uma informação.

Diversos estudos foram realizados a fim de definir conhecimento e de compreender e explicar seu processo de aquisição e raciocínio. Desses estudos, os mais importantes e atuais enquadram-se dentro das áreas de sociologia, psicologia e cognição (WIVES, 2004). Nessas áreas, o conhecimento é compreendido como sendo a forma com que a pessoa percebe o mundo. Por estar em constante interação com o meio, o conhecimento de uma pessoa muda com o tempo. Assim, o conhecimento de uma pessoa em determinado momento é denominado *estado de conhecimento* (MIZZARO, 1996).

O conhecimento é o que se aprende da informação e, que possa ser utilizado para a compreensão de novas situações que ocorrem no mundo real (HERSH, 1996).

### 2.2.5 O significado

Na área de informações, esse é o conceito mais difícil de ser definido. Sugerir que as

palavras são simples símbolos para descrever as coisas do mundo é ingênuo e uma simplificação grosseira. As palavras são traiçoeiras.

> *"Nenhuma palavra possui exatamente o mesmo sentido duas vezes (HAYAKAWA, 1939)".*

O real significado de uma palavra não será claro até que se descubra o contexto na qual está inserida. E o contexto é um componente tão sutil quanto um trocadilho, uma palavra ambígua, uma piada. Além disso, o significado depende de quem fala, de quem escuta, do nível de conhecimento e da experiência para interpretação e talvez até da situação geográfica.

Muitas teorias semânticas ainda são controversas a respeito da definição de significado, e de sentido. Essas definições são vistas de formas diferentes pelas diferentes disciplinas como Filosofia, Ciência Cognitiva e Ciências da Informação.

A base da teoria semântica, a teoria dos signos, a semiótica, de uma forma ou outra, recorrem tradicionalmente a um modelo conhecido como o triângulo semiótico para explicar os processos perceptivos, cognitivos e pragmáticos ligados ao uso de signos. Os três pólos do triângulo semiótico são o signo, o significado e o objeto real ao qual ambos se referem. E essa relação triádica que domina o tema remonta desde a Antigüidade Grega (BLIKSTEIN, 1990; ECO, 1996).

Outros autores formularam outros modelos para explicar processos perceptivos, cognitivos e pragmáticos. Entre eles, pode-se citar Frege e Jakobson. O próprio Blikstein complementa o modelo de Heger ("conceito" e "coisa") através de seu modelo em que se funde "signo" e "significado" (referência) e renomeia como "língua", pois, a língua influencia a práxis social que, por sua vez, determina o aparelho de percepção e cognição, que estrutura a realidade amorfa e é alimentado e alterado por ela ao mesmo tempo – diria um sistema realimentado de "malha fechada". Por último, o aparelho cognitivo reformula, através do referente, o sistema lingüístico (ECO, 1996). A figura 3 apresenta o modelo citado com elementos extralingüísticos transcendendo qualitativamente o triângulo tradicional.

O Gráfico de Blikstein mostra uma preocupação com o perceptivo-cognitivo triângulo semiótico. O signo, como momento (sempre em crise) do processo de simiose, é o instrumento através do qual o próprio sujeito se constrói e se desconstrói constantemente. A ciência dos signos é a ciência de como se constitui historicamente o sujeito (ECO, 1996).

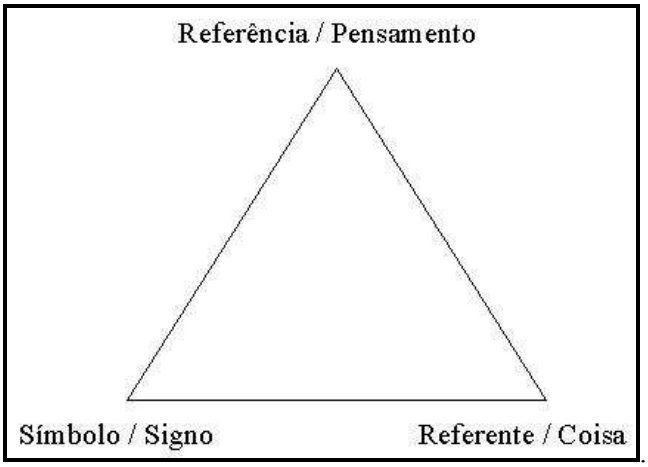



Figura 2: Triângulo Semiótico de OGDEN e RICHARDS (OGDEN e RICHARDS, 1956)

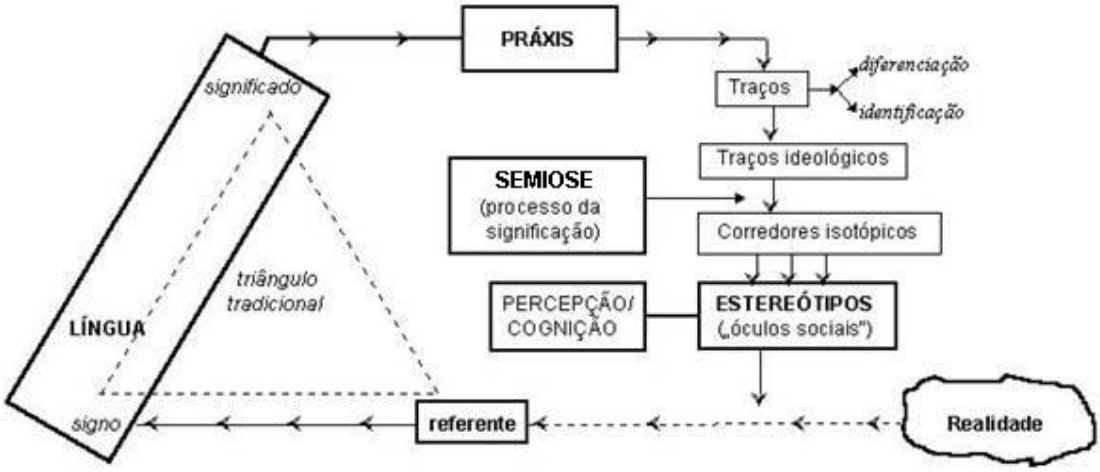



Figura 3: Gráfico do modelo de Blikstein (adaptado de BLIKSTEIN, 1990)

Neste trabalho, o sentido de uma expressão lingüística é definido pela construção mental associada às entidades do mundo real ou abstrato, de acordo com o Triângulo Semiótico de Ogden e Richards (OGDEN e RICHARDS, 1956).

## 2.3 SISTEMA DE INFORMAÇÕES

Um Sistema de Informação pode ser uma biblioteca, pública ou especializada; um centro de documentação de uma empresa; um arquivo, um museu ou um banco de dados. Seja qual for a sua denominação original, um Sistema de Informação tem por função coletar, tratar e disseminar a informação produzida pela sociedade na qual está inserido, garantindo, assim,

o acesso à cultura por parte de seus membros e possibilitando a sua continuidade (LIMA, 1998).

Buckland (BUCKLAND, 1991) define Sistemas de Informação como quaisquer unidades que coletem, tratem, organizem e disponibilizem "coisas" potencialmente informativas.

## 2.4 RECUPERAÇÃO DE INFORMAÇÃO

### 2.4.1 Recuperação de informação como processo iterativo

RI é um processo de comunicação. Um sistema de Recuperação de Informação consiste de uma base de dados – onde são armazenados e disponibilizados os dados - e um *software* para processar entradas e saídas. Na terminologia convencional de base de dados, os itens na base de dados são chamados de registros. Na terminologia de Recuperação de Informação, entretanto, os registros são chamados de documentos e, portanto, pode-se chamar de Base de Dados de Documentos para a área de RI (HERSH, 1996).

A Recuperação é um processo de interação com um SRI no sentido de obter documentos – não necessariamente relevantes. Um usuário interage com o sistema através de uma *necessidade de informação*. BELKIN (BELKIN e CROFT, 1992), descreve essa necessidade como sendo "Estado Anômalo do Conhecimento" (*anomalous state of knowledge – or ASK*). O usuário, especialista ou não, formula uma necessidade de informação através de uma expressão de busca (*query*), a qual normalmente consiste de termos de um ou mais vocabulários indexados que podem ser conectados por operadores *booleanos* (AND, OR ou NOT). Após sua submissão, o sistema processa a expressão de busca e retorna o os documentos encontrados para o usuário.

### 2.4.2 Sistema de Recuperação de Informação e Gerenciador de Banco de Dados

Uma outra forma de entender sistemas computacionais é comparar as aplicações que são executadas. Um SRI não é o mesmo que um sistema de gerenciamento de banco de dados (SGDB). Um sistema típico de SGBD disponibiliza bases de dados altamente estruturados. Nesse sistema, a resposta a uma pergunta existe ou não existe na base de dados, como, por exemplo, o número único de prontuário. Num sistema de Recuperação de Informações, a

resposta para uma questão específica talvez possa existir ou talvez não exista, e ainda, se existir, pode não ser fácil encontrá-la (HERSH, 1996).

Outra diferença entre um Sistema de Recuperação de Informações e um SGDB é o registro na base de dados. Num SGDB, o registro possui um ou mais campos, com características previamente determinadas, nas quais cada uma consiste num tipo específico de informação. Por exemplo, uma base de dados de pacientes poderá constar de campos para registrar, além dos dados essenciais do paciente, outros relativos ao histórico enquanto paciente na instituição, tais como data de entrada, prescrição de remédios, data da alta, e informações relativas ao diagnóstico, entre outros. O registro no Sistema poderá ter somente dois campos, por exemplo, um para registrar título e outro para registrar um texto livre; ou ainda, como se pode verificar em algumas bases de dados bibliográficos especializados, que possuem inúmeros campos para títulos, *abstracts*, tipo de publicação, etc... Alguns desses campos, com tamanho fixo e tipo de dados determinados, podem ser considerados semelhantes aos utilizados num sistema de gerenciamento de banco de dados – *data management system* (DBMS); porém, outros campos contêm textos de tamanhos variados.

Outra diferença entre os dois tipos de sistema está na forma como os dados são indexados. Além de discriminar os descritores para representar o conteúdo de um registro ou campo, a outra proposta de indexação permite um rápido acesso aos registros ou aos documentos baseados no seu conteúdo. Num SGBD, pode-se ter uma ou mais chaves, onde cada uma é derivada do conteúdo inteiro de um simples campo, tal como, o número único de um prontuário médico. Num sistema de Recuperação de Informação, por outro lado, o processo de indexação poderá considerar o termo completo, que poderá conter mais de uma palavra (palavras compostas); parte do termo (como se fosse um *stems* ou radical); e ainda, desconsiderar termos completos, normalmente considerados como *stopwords* (em geral, termos sem peso semântico ao processo).

Nesse processo de indexação, podem ser utilizados procedimentos complicados, como técnicas que permitam, por exemplo, mapear termos sinônimos ou textos e vários campos para os termos de um vocabulário controlado.

Todavia, o limiar que divide as diferenças entre SRI e SGDB está cada vez mais tênue, pois alguns SGDB modernos geralmente incluem funcionalidades de RI.

### 2.4.3 Aspectos de Sistemas de Recuperação de Informações

Outra forma de entender os Sistemas de Recuperação de Informações é analisar os

processos utilizados em Recuperação de Informações. Existem várias facetas utilizadas no processo de Recuperação de Informações e serão apresentados três aspectos que descrevem de maneira abrangente o processo de funcionamento e sua interação com o usuário final.

2.4.3.1 O modelo de SRI de Meadow

A figura 4 mostra o ciclo e fluxo de informações que é utilizado num sistema de recuperação de informações interagindo com o usuário (MEADOW, BOYCE e KRAFT, 1992).

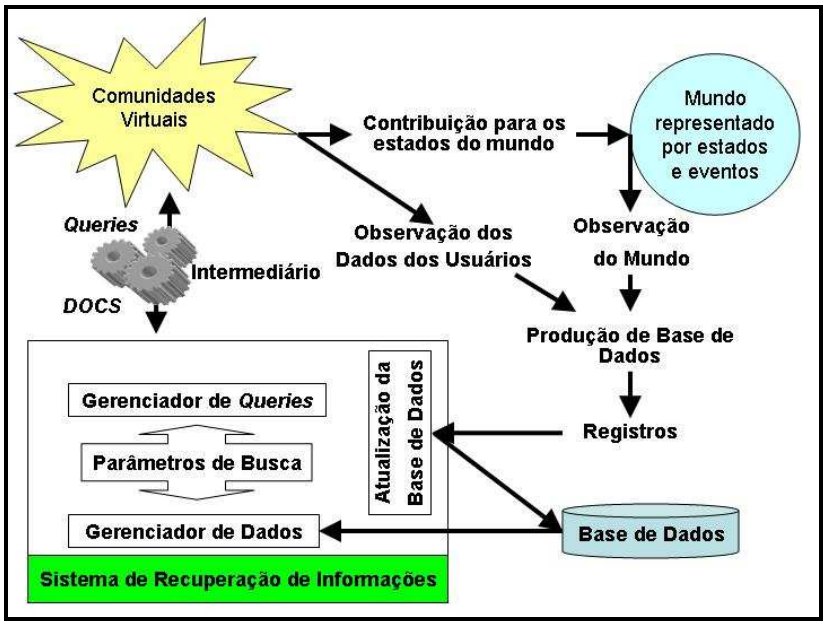



Figura 4: Modelo de fluxo de informações no mundo real (HERSH, 1996).

O sistema é cíclico com a informação, geralmente fluindo pelo lado direito nesse diagrama. A informação sobre o mundo real vem de uma comunidade de usuários, que também representam os próprios criadores da informação, utilizada nesse processo para afetar certos conceitos do mundo real e, conseqüentemente, a concepção de mundo dos próprios usuários.

Começando da criação da informação, verifica-se que eventos que ocorrem no mundo real são transcritos em forma de periódicos, livros, jornais e outros tipos de publicação. Geralmente, essas bases são construídas e organizadas no banco como registros para serem utilizadas em Sistemas de Recuperação de Informações. Então, os usuários podem, através

deste sistema, formular questões ao banco de dados e recuperar informações dos registros. As informações recuperadas podem ser utilizadas pelo usuário para gerar novas contribuições ao mundo. Além disso, através da observação, o usuário pode realimentar o banco pela adição de novas informações ou simplesmente melhorar a qualidade do banco de dados através da análise da informação recuperada pela correção de erros encontrados.

2.4.3.2 O modelo de SRI de Salton

No modelo de Salton (SALTON, 1983), o cerne de um SRI, conforme mostra a figura 5, possui foco voltado de como os itens dos registros do banco podem ser combinados com a expressão de busca do usuário. Em particular, do ponto de vista da área de recuperação de informações, os registros, ou seja, os documentos (DOCS) de um banco de dados são descritos utilizando um conjunto de descritores, nomeadamente linguagem indexada (LANG). Nesse processo de indexação, os descritores de uma linguagem indexada são mapeados para cada termo do documento - em alguns sistemas pode-se ter mais de uma linguagem indexada (HERSH, 1996). Na recuperação ou processo de formulação da busca, o usuário entra com uma expressão de busca no sistema a qual é transformada na linguagem indexada – que pode ser uma linguagem independente do usuário. Então, documentos candidatos são devolvidos ao usuário, após um processo de medida de similaridade entre a expressão de busca do usuário e documentos (que não necessariamente estejam num banco de dados).

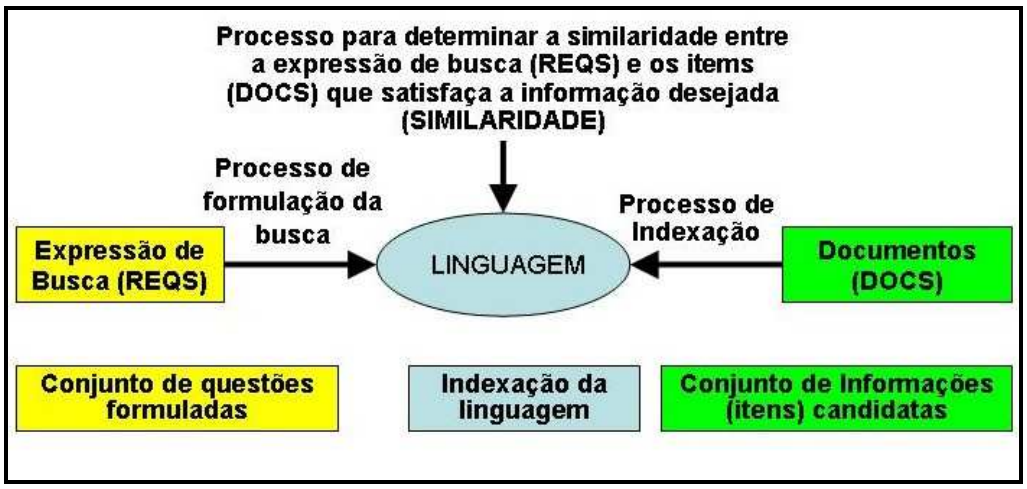



Figura 5: Modelo de Salton de um Sistema de Recuperação de Informações (HERSH, 1996).

2.4.3.3 O modelo de SRI de Marchionini

A figura 6 mostra o cenário função busca-informação do ponto de vista do usuário (MARCHIONINI, 1992). O componente central é a definição do problema pelo usuário (ou a necessidade da informação). Uma vez definida, o usuário seleciona a fonte a ser pesquisada e formula a questão. O usuário realiza a busca, examina os documentos entregues pelo sistema, e extrai a informação do conjunto.

Nesse modelo, o processo de formulação da pergunta pode ser interativo e o usuário poderá reformular as *queries*. Às vezes, os resultados obtidos podem levar o usuário a uma nova necessidade de informação; ou ainda, forçar a mudar a estratégia de busca.

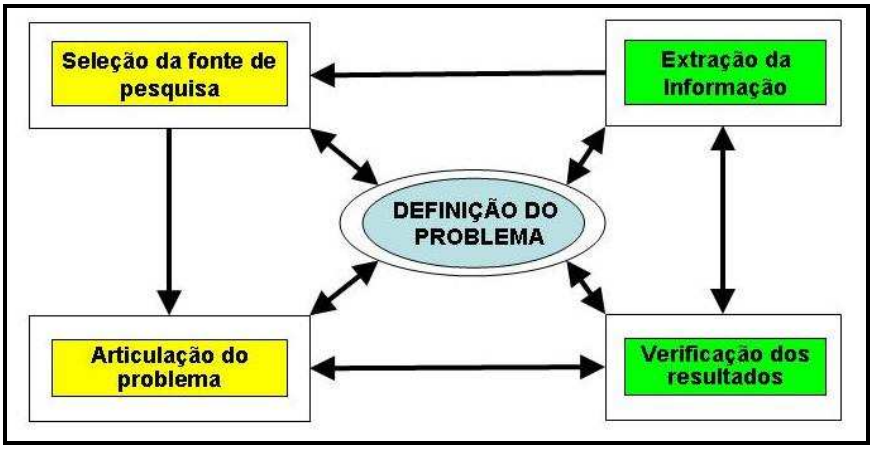



Figura 6: Modelo de Marchionini de um sistema Recuperação de Informações (HERSH, 1996).

Nesse processo, os resultados poderão levar o usuário a mudar a estratégia de formulação das *queries.*

## 2.5 AVALIAÇÃO DE SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO

Existe várias razões do porquê a avaliação de um SRI é importante. É um processo complexo, que por vezes envolve muita infra-estrutura de *software* e *hardware*. Mas a principal razão, entre outras, é determinar o quanto um SRI, desenvolvido num certo domínio, é eficaz em responder a uma necessidade de informação de um usuário (HERSH, 1996).

### 2.5.1 A Conferência TREC

A TREC (*Text Retrieval Conference*) é uma conferência que trata de avaliação de RI sob o ponto de vista de sistemas. Nela, comparam-se as diversas técnicas utilizadas pelos grupos participantes. Para cada tarefa existe uma base de documentos com cerca de 2 GB de texto e 50 consultas que informam o que é a informação procurada e o que constitui um documento relevante. Esse sistema também é alvo de críticas por realizar as avaliações em um ambiente de laboratório. Na realidade, sistemas de avaliação baseadas em julgamentos de relevâncias serão sempre criticados, pois o julgamento em si é subjetivo.

Do ponto de vista do usuário, não existe uma metodologia de avaliação padrão. Para avaliar o comportamento, necessidades e satisfação dos usuários, os métodos incluem: entrevistas, observações, experimentos e pesquisa (AIRES, 2002). Este tipo de avaliação é caro, demorado, mas tem a vantagem de refletir melhor a real necessidade dos usuários.

### 2.5.2 A Medida de Precisão e Revocação

A avaliação mais comum em RI é realizada sob o ponto de vista de dois parâmetros que é a Precisão (P) e a Revocação (R) – *Precision Recall*. Outras medidas utilizadas são a medida F (*F-Measure*), a medida E e o *Fallout* (RIJSBERGEN, 1979). Mas há controvérsias sobre a confiabilidade de tais medidas, independente da escolha do tipo de medida a ser utilizada. Uma questão discutida, por exemplo, é a relevância das pequenas diferenças sobre o sucesso da busca realizada por meio de um usuário (GWIZDKA e CHIGNELL, 1999).

Em muitas situações, normalmente se tem uma seleção de documentos (falsos positivos – *fp*) relevantes (por exemplo, de possíveis documentos relevantes ou de sentenças nas quais as palavras possuem um certo sentido, por exemplo) de uma de uma coleção muito grande (negativos verdadeiros - *tn*) que possui um conjunto de respostas (amostra de falsos negativos - *fn*). Além disso, há documentos relevantes encontrados nesse conjunto de respostas (positivos verdadeiros - *tp*). Essa situação pode ser esquematizada na figura 7, na qual pode-se agrupar a amostra e a seleção como variáveis randômicas e sua distribuição pode ser expressa em termos de duas variáveis como uma matriz contingência, conforme descrito na tabela 1.

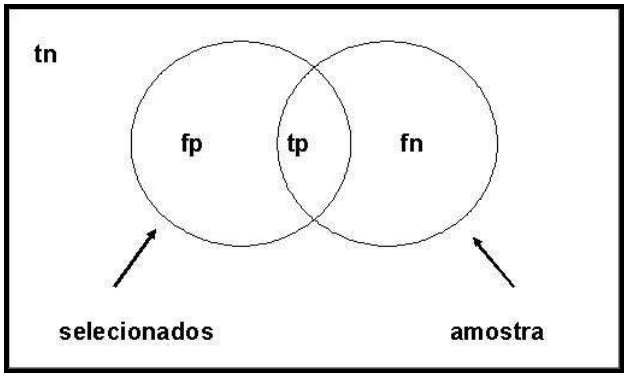



Figura 7: Diagrama motivacional da medida de precisão e revocação (MANNING e SCHÜTZE, 1999).

A figura 7 mostra as áreas representando os positivos verdadeiros e negativos verdadeiros (*tp, tn*), os falsos positivos e os falsos negativos (*fp, fn*) são apresentados em termos de amostra selecionados e itens selecionados da amostra.

Tabela 1: Precisão e Revocação – variáveis randômicas e sua distribuição em termos de duas variáveis como matrix de contigência 2 x 2.

| | Atual | |
|---|---|---|
| Sistema | Amostra | ¬ Amostra |
| Selecionado | *tp* | *fp* |
| ¬ Selecionado | *fn* | *tn* |

A tabela 1 mostra as freqüências ou a quantidade de cada item em cada região do espaço representado na figura 7. Os casos assinalados por tp (positivos verdadeiros) e tn (negativos verdadeiros) são os casos corretos para o sistema. O caso de seleção errada, assinalada para fp é chamado de falsos positivos, falsos aceite ou erro do tipo II. O caso fn representa os falsos negativos, falsa rejeição ou erro do tipo I (MANNING e SCHÜTZE, 1999).

A precisão é definida como a medida proporcional dos itens selecionados para os casos corretos do sistema (equação 2). A revocação é definida como a proporção de itens do da amostra selecionada pelo sistema (equação 3).

$$P = \frac{tp}{tp + fp} \tag{2}$$

$$R = \frac{tp}{tp + fn} \tag{3}$$

Em muitas aplicações, somente os parâmetros de precisão e revocação não fazem muito sentido para a área de Processamento de Linguagem Natural. O parâmetro precisão é calculado para diversos níveis de revocação. Dependendo do que se quer, o interessante é ter um valor médio que envolva os dois parâmetros. Assim, uma medida preferida que combina ambos é a Medida F (*F-measure*) (SABATER e SIERRA, 2005).

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha)\frac{1}{R}} \tag{4}$$

onde P é a precisão , R a revocação e α determina o peso entre precisão e revocação, normalmente em torno de 0,5 para o peso entre os parâmetros; e com esse valor, pode-se simplificá-la conforme indica a equação 5.

$$F = \frac{2PR}{(R + P)} \tag{5}$$

### 2.5.3 A Coleção de Teste OHSUMED

A coleção OHSUMED foi criada para dar suporte aos pesquisadores na área de RI em saúde. De acordo com Hersh (HERSH, BUCKLEY, LEONE e HICKAM, 1994), a coleção OHSUMED é um conjunto de 348.566 documentos médicos clínicos extraídos da MEDLINE (de um total de mais de 7 milhões de documentos) que cobre todas as referências dos 270 jornais de um período de cinco anos (1987-1991). Seu tamanho é de aproximadamente 400 MB. A coleção inclui documentos escritos em inglês que são estruturados em 7 campos: identificador, título, fonte, autores, termos MeSH, tipo de publicação e resumo.

A coleção OHSUMED inclui um conjunto de 106 consultas textuais escritas em inglês, cujo conjunto ideal de respostas, julgamento de relevância, foram identificadas por especialistas em saúde. Existe um total de 16.140 pares de *queries* e documentos relacionados pelo julgamento de relevância.

Nesse trabalho, utilizou-se, para a verificação dos resultados da proposta, através da técnica de precisão e revocação, um subconjunto de 233.445 (67%) documentos que contêm obrigatoriamente o campo resumo e as 106 *queries* para plotar a *baseline* como referência para outras línguas como medida de desempenho.

## 2.6 VOCABULÁRIO CONTROLADO

Segundo Miller (MILLER, 1997), tesauro é definido como um "modelo léxico-semântico de realidades conceituais ou suas constituintes expressas na forma de um sistema de termos e suas relações, que oferece acesso via diferentes aspectos e é usado como ferramenta no processamento e busca de uma unidade de recuperação de informação".

O tesauro no campo da informação e documentação é uma lista organizada de conceitos compilados que serve para indexar e recuperar documentos de um certo domínio. A idéia não se resume somente à definição de termos na construção do léxico, mas também tratar dos relacionamentos entre eles (HUGE, 1999). São relações do tipo sinonímia, hiperônimos (carro, automóvel), hipônimos (automóvel, carro), relação parte-de (mão, dedos), antônimos (aceleração, desaceleração) e compatibilidade (carro, volante). O tesauro possui diferentes funções no campo da informação e documentação. Durante a produção de documentação, ele pode ser utilizado para normalizar o vocabulário contido nos documentos. Ele também pode ser utilizado para a construção de uma representação de documentos para uma abordagem de recuperação (BAEZA-YATES e RIBEIRO-NETO, 1999).

De acordo com FOSKETT (FOSKETT, 1997), a idéia principal de se utilizar um tesauro é prover um vocabulário controlado de referência a um sistema de recuperação de informações – indexação e busca.

Na área de recuperação de informações, a indexação é definida como uma forma de mapear assuntos dos documentos. Existem duas razões para indexar uma coleção de documentos. A primeira é representar os assuntos de cada documento para que possam ser recuperados por um usuário; e a segunda, de organizar os diversos assuntos de forma que programas de computador possam localizar rapidamente os documentos com assuntos referentes a um determinado conceito (HERSH, 1996).

As abordagens para a construção de tesauro são basicamente duas: manual e automática. Normalmente, torna-se necessário e mesmo obrigatório construir manualmente o tesauro devido à complexidade de relacionamentos entre conceitos, as ambigüidades

semânticas e o próprio dinamismo inerente a cada língua. A construção demanda muito tempo e sua manutenção é complexa (SANCHES, dez/1997).

## 2.7 LINGÜÍSTICA DE *CORPUS*

A Lingüística de *Corpus* é a área que utiliza a observação de dados estatísticos e probabilísticos advindo do processamento de corpus de texto com o objetivo de levantar características lingüísticas (SEATON, 1995).

A existência de uma coletânea de dados lingüísticos naturais, legíveis por computador é central à Lingüística de Corpus atual. Porém, nem todo conjunto de dados é considerado um *corpus* (SARDINHA, 2004). Suas principais definições são:

(1) arquivo: depósito de textos sem organização prévia;
(2) biblioteca eletrônica: coleção que segue alguns critérios de seleção;
(3) *corpus*: uma parte da biblioteca eletrônica, construído a partir de um projeto explícito, com objetivos específicos;
(4) *subcorpus*: uma parte de um *corpus* que pode ser fixa ou mutável (dinâmica, isto é, flexível durante a análise) (ATKINS e OSTLER, 1992).

Definições de *corpus* proliferam-se na literatura, tal como a apresentada por Sinclar (SINCLAIR, 1995): “uma coletânea de textos naturais, escolhidos para caracterizar um estado ou variedade da linguagem”.

Textos Naturais são aqueles que existem na linguagem e que não foram criados com o propósito de figurarem no *corpus*. Além disso, amplia-se a idéia de natural para incluir somente aqueles textos produzidos por seres humanos. Dessa forma, está excluída a produção provinda de programas de geração de textos. Um problema com essa definição é que não deixa claro o propósito da criação do *corpus*. Por isso, deve ser incorporada a complementação: “*corpus* é um corpo de linguagem natural (autêntica) que pode ser usado como base para pesquisa lingüística” (SINCLAIR, 1995).

Assim, embora os textos devam ser naturais (autênticos e independentes do corpus), o *corpus* em si é artificial, um objeto selecionado com critérios previamente definidos, com fins específicos de pesquisa. Esses dois posicionamentos estão presentes: “*corpus* é uma coletânea de porções de linguagem que são selecionados e organizados de acordo com critérios lingüísticos explícitos, a fim de serem usadas como uma amostra da linguagem” (PERCY e MEYER, 1996).

A definição a seguir faz menção à extensão do *corpus*: "uma coletânea grande e criteriosa de textos naturais" (SARDINHA, 2004). Por criteriosa entende-se que deva refletir variedade o mais fielmente possível; ou seja, para um *corpus* geral de uma língua, deve-se incluir a maior quantidade de ocorrência de palavras possíveis no domínio em questão. Se por outro lado, for um *corpus* específico, deve-se ser o mais seletivo possível na escolha de exemplares, para que os mesmos reflitam de fato a variedade escolhida, ou seja, para que não haja vieses ou contaminações.

Incorporando as características já mencionadas nas anteriores tem-se que "*corpus* é um conjunto de dados lingüísticos (pertencente ao uso oral ou escrito da língua, ou ambos), sistematizados segundo determinados critérios, suficientemente extenso em amplitude e profundidade, de maneira que sejam representativos da totalidade ou do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise" (SANCHES, dez/1997).

A linguagem é um sistema probabilístico (HALLIDAY, 1991), no qual certos traços são mais freqüentes que outros. Pode-se diferenciar as palavras entre aquelas de maior freqüência e as de menor freqüência, sendo que a diferença entre elas é relativa. Assim, algumas palavras têm freqüência de ocorrência muito rara e, para que haja probabilidade de ocorrência no *corpus*, é necessário incorporar uma grande quantidade de palavras. Portanto, quanto maior a quantidade de palavras, maior a probabilidade de aparecerem palavras de baixa freqüência.

No caso dos sentidos das palavras, pode-se também distinguir entre os sentidos mais freqüentes e os menos freqüentes das entidades lexicais. Assim, mesmo palavras de alta freqüência têm sentidos raros (por exemplo, "serviço" entendido como saque no jogo de tênis) que terão maior probabilidade de ocorrer quanto maior for o *corpus*.

O *corpus* é uma amostra de uma linguagem como um todo, cuja dimensão não se conhece. Desse modo, não se pode estabelecer qual seria o tamanho ideal da amostra para que represente essa população. Uma salvaguarda é tornar a amostra o maior possível (SINCLAIR, 1995), a fim de que ela se aproxime ao máximo da população da qual deriva, tornando-se, portanto, mais representativa. Para que ela seja representativa, é necessário conhecer a população da qual ela provém.

A freqüência em si não é suficiente, porque mesmo palavras de alta freqüência possuem vários sentidos. Assim, uma freqüência alta pode esconder vários sentidos, que separados teriam baixa freqüência. Para que seja representativo, um *corpus* deve conter o

maior número possível de sentidos de cada forma. Por exemplo, a forma “como” pode significar a preposição ou a primeira pessoa do singular do verbo comer no presente do indicativo. Essa forma é comum na língua portuguesa, ocorrendo aproximadamente 531 vezes por milhão (SARDINHA, 2004).

Um modo de atingir a representatividade total de um *corpus* é incluir nele toda a linguagem. Como é impossível para um idioma inteiro, a possibilidade mais próxima é restringir o conteúdo a um autor ou assunto apenas, por exemplo.

## 2.8 PROBLEMAS DE CODIFICAÇÃO

Codificação diz respeito a uma representação de um símbolo baseada num modelo de distribuição probabilística. A idéia geral é que um sistema com código curto possa representar os símbolos mais comuns enquanto que os códigos mais longos possam representar os símbolos raros, isto porque se considera um fator muito importante: a velocidade de processamento dos símbolos codificados (WITTEN, 1994).

Durante algum tempo muitos sistemas de computador operavam somente com a escrita pertencente ao sistema ASCII (*American Standard Coding for Infomation Interchange*) - padronizado em 1986 (ANSI X3.4, RFC 20, ISSO/IEC 646:1991, ECMA-6) pelo *American National Standards Institute*, ou seja, utilizavam os mesmos caracteres utilizados no inglês. Para operar com a escrita de outros idiomas, era necessário adotar um padrão diferente que não eram intercambiáveis entre si como, por exemplo, o JUNET para o Japonês, ou o ASCII estendido para o idioma latino. Embora o ASCII fosse suficiente para comunicação em inglês moderno, em outras línguas como as européias e latinas que incluem caracteres acentuados, as coisas não foram tão fáceis. Os padrões ISO 8859 foram desenvolvidos para satisfazer a essas necessidades (ABAITUA, 2002).

O ASCII utiliza sete bits, ou seja, utiliza padrões de dígitos representáveis por sete dígitos binários, o que fornece um alcance de 0 a 127 em decimais. Isto inclui 32 caracteres de controle não visíveis, a maior parte entre 0 e 31 e com o caractere de controle final, DEL ou delete em 127. Os caracteres de 32 a 126 são caracteres visíveis: um espaço, marcas de pontuação, letras latinas e números (WITTEN, 1994). Essa miscelânea de padrões levou a muita confusão, e também a uma quase total falta de capacidade para comunicação multilíngüe, especialmente em diferentes alfabetos. Mas a internet propiciou a implementação de uma solução mais homogênea.

Um protocolo é um conjunto de regras que governa um processo. *Hypertext Transfer Protocol* - HTTP é o protocolo base para *web-browsers* e foi projetado com vistas a transferência de arquivos (YERGEAU, ADAMS e DUERST, 1997). Esse sistema foi projetado para a transmissão de meta-informação sensível ao idioma. Um protocolo efetivo dever ter sua especificação documentada. As especificações estabelecem o formato exato de como os dados devem transitar. A especificação de protocolos para internet é chamada de *Request for Comments* - RFC (HEATON, 2002). Normalmente, o RFC tem um número associado a uma norma específica. A norma RFC 2068 da versão HTTP 1.1 contempla a codificação de caracteres e a negociação "lingüística" cliente-servidor. De acordo com a RFC 2068, a codificação dos caracteres se indica mediante um parâmetro no campo cabeçalho (*header*) do protocolo. Um arquivo em japonês codificado com JUNET, por exemplo, conterá no cabeçalho o protocolo com os atributos "Content-type: text/htm; charset=isso-2022-JP. O cliente poderá indicar a preferência por uma determinada codificação (Accept-Charset) e o idioma (Accept-Language) (ABAITUA, 2002).

A norma RFC 2070 relaciona questões com conjuntos de caracteres. A norma RFC 1886, adotada nas primeiras versões da linguagem de marcação HTML, restringe o conjunto de caracteres no padrão ISO-8859-1 ou ISO-Latin-1, que só serve para línguas com o alfabeto latino. Esse padrão utiliza 8 bits de forma que permite representar no máximo de 256 caracteres. A norma posterior, RFC 2070, incrementou propriedades ao HTML de forma a suportar documentos em outros idiomas. A ISO-Latin tem sido substituída pela ISO-10646 de 1993, mais conhecida como UCS (Universal Character Set), e que coincide com a norma UNICODE 1.1. UNICODE é um sistema de 16 bits e com isso é capaz de representar todos os sistemas de escrita no mundo. O Unicode livra-se da limitação de um único bit tradicional dos conjuntos de caracteres. Ele usa 17 "planos" de 65.536 pontos de código para descrever um máximo de 1,114,112 caracteres. O Unicode foi mapeado de diversas maneiras, mas os dois mais comuns são o **UTF** *(Unicode Transformation Format)* e **UCS** *(Universal Character Set).* O número após UTF indica o número de bits em uma unidade, enquanto o número após UCS indica o número de bytes. UTF-8 tornou-se o meio mais comum de intercâmbio de texto em *Unicode* como resultado de sua natureza limpa de oito bits. O UTF-8 é uma codificação de caracteres de tamanhos variáveis, o que neste exemplo significa que usa de 1 a 4 bytes por símbolo. O primeiro byte de UTF-8 é usado para codificar ASCII, dando ao conjunto de caracteres compatibilidade com ASCII. UTF-8 significa que ASCII e caracteres latinos são intercambiáveis com pouco aumento no tamanho dos dados, porque somente o primeiro bit é usado (ABAITUA, 2002). UTF-8 permite que você trabalhe em um ambiente multilíngüe e

internacionalmente aceito que atende a padrões, com uma redundância de dados comparativamente baixa. UTF-8 é o modo preferível de se transmitir caracteres não-ASCII através da Internet, através de E-Mail, IRC ou qualquer outro meio. Pelas suas características, UTF-8 é considerado um excesso para comunicação via internet através de E-mail, IRC ou serviços semelhantes.

## 2.9 ANOTAÇÕES E LINGUAGENS DE MARCAÇÃO

Anotações podem ser utilizadas para etiquetar os termos de um *corpus* com o intuito de melhor organizar os itens do próprio *corpus*. Um corpus etiqueta possibilita uma busca mais precisa por certos tipos de informação contida nele. Na prática, a maioria dos corpora possui algum tipo de anotação.

Existem basicamente dois tipos de marcação: aquelas relacionadas à identificação do texto, como, por exemplo, o título e o autor; e aqueles que se aplicam em parte do conteúdo (usualmente refere-se à palavra ou um grupo de palavras), como por exemplo, *part-of-speech* ou referências anafóricas.

Usualmente, um *corpus* contém um texto puro, sem formatação, a menos que tenha sido obtido de uma fonte de publicação (como artigos de revistas eletrônicas ou *newspapers*) ou de coleções estruturadas (*databases*). Nesses casos, essas fontes possuem informações extras anexadas no início do documento como cabeçalho (*header*), as quais descrevem informações do tipo, título, a data de publicação, área de concentração e assim por diante. Para a realização de processamento, torna-se necessário separar as informações extras do conteúdo textual, e isso é relativamente fácil (GALLE, JAKOBS, KESTEN *et al.*, 1992).

Arquivos em formato puro de texto, sem formatação, são os mais simples e limitados. Não é possível representar caracteres não ASCII, como letras acentuadas e *umlauts,* o que impõe sérias restrições para trabalhar com *corpora* que não sejam no inglês. Após anos desenvolvendo linguagem de marcação (*mark-up language*) idiossincrática, um formato foi estabelecido: o padrão SGML e a sua versão simplificada, XML. Sua especificação formal descreve como codificar um texto e representar a informação (MASON, 2000).

Linguagem de marcação é um conjunto de códigos aplicados a um texto ou dados com a finalidade de adicionar informações particulares sobre esses textos ou dados, ou sobre trechos específicos. As marcações são feitas com etiquetadores (*tags*). O etiquetador serve para inserir automaticamente no *corpus*, códigos que indicam a classe gramatical de cada

palavra ou estruturas que definem instruções, tendo uma marca de início e outra de fim. A etiquetagem pode ser automática ou semi-automática (interativa). Há vários tipos:

(1) Morfossintática (*part-of-speech* ou *pos*): marcação da classe gramatical (substantivo, verbo, adjetivo, etc...) de cada palavra. Também chamado de morfológica, é a mais comum.
(2) Sintática (*parsing*): identificação da estrutura sintática (sintagma nominal, verbal, etc.) de cadas frase.
(3) Semântica (*semantic*): definição do sentido ou da categoria semântica da cada palavra (por exemplo, casas = moradia, martelo=ferramenta).
(4) Discursiva (*discourse*): marcação de características como referentes anafóricos, tópicos ou marcadores discursivos (SARDINHA, 2004).

O SGML (*Standard Generalized Markup Language*) é uma linguagem de marcação criada no final da década de 1960 com o objetivo de construir um sistema portável; ou seja, que fosse independente de sistema operacional, formatos de arquivos, etc., de tal modo que pudesse compartilhar informações para a realização de algum processamento. Desta forma, definiu-se um sistema de Marcação Generalizada (*Generalized Markup*), em que os nomes das marcações seriam definidos pelo usuário, permitindo customizar um padrão de detalhamento dos dados (MASON, 2000). Esse sistema possui dois objetivos básicos:

(1) Descrever a estrutura do documento e outros atributos que lhe são importantes. Assim, o processamento das informações pode ser automatizado, já que não é necessário especificar o processamento a ser feito. Isto torna o documento autodescritivo;
(2) Garantir o processamento através de uma marcação rígida a fim de evitar falha devido à má formatação por um usuário ou por um software na construção de um documento.

A marcação generalizada não restringe documentos a uma única aplicação, estilo de formatação ou sistema de processamento. SGML foi, portanto, uma evolução na forma de compartilhar informação. Com o advento da Internet, um ambiente tão heterogêneo[5], esse tipo de linguagem logo se tornou um padrão internacional muito utilizado. E, assim, o SGML adquiriu três características básicas: marcações descritivas, as marcações podem ser tipadas e, independência de plataforma.

Com a marcação descritiva, um documento pode ser processado em partes, e também em diferentes softwares.

[5] Do ponto de vista dos sistemas operacionais, tecnologias, linguagens de programação e plataforma, a Internet é heterogênea.

O SGML traz o conceito de tipo de definição de documento, os DTD (*Document Type Definition*). Para que os softwares não carreguem consigo as informações dos tipos de dados de um documento, tornando-os mais específicos e diminuindo a aplicabilidade do padrão SGML nas diferentes plataformas, criou-se os DTDs para detalhar os tipos que um documento comporta. Os DTDs fornecem meios para definir os tipos de dados. Se tal especificação não for definida, provavelmente um software irá gerar um erro por não saber como tratar determinados tipos de dados; se são *strings*, data, números, etc.

A característica básica do SGML é assegurar que os dados sejam mantidos, não importando em que plataforma de software ou hardware.

O XML (*Extensible Markup Language*) é um padrão para publicação, combinação e intercâmbio de documentos multimídia, desenvolvido pelo consórcio W3C (*World Wide Web Consortium*). O XML utiliza o padrão de codificação UNICODE (ABAITUA, 2002).

A definição da linguagem XML consiste em padrão de marcação com um conjunto de "*tags*", onde contém informações estruturadas, ou seja, documentos que contêm uma estrutura clara e precisa da informação que é armazenada em seu conteúdo (OLIVEIRA, 2002). A capacidade de descrever dados é chamado de "*self-describe data*".

No Sistema *Morphosaurus*, o tesauro é exportado para o padrão XML, com base num arquivo DTD, para ser utilizado pelo módulo de segmentação do sistema. A figura 8 ilustra parte do conteúdo do referido arquivo. Esse exemplo mostra três *subwords* na língua portuguesa, delimitadas pelas etiquetas </lex>...</lex>, que contém os elementos que caracterizam as *subwords* delimitadas pelas etiquetas <mid>...</mid>, <str>...</str>, <t>...</t> e <l>...</l>.

Nesse arquivo gerado as etiquetas apresentam um tipo de informação:

(1) "<lex>" e "</lex>" que determina o início e o final de cada lexema;
(2) "<mid>" e "</mid>" que representa o conceito de forma multilíngüe. É a linguagem "artificial" do *Morphosaurus*;
(3) "<str>" e "</str>" que determina um termo lexical;
(4) "<t>" e "</t>" que determina qual é o tipo do lexema, por pedido do sistema *Morphosaurus* é representada por siglas, ST – Radical, PF – Prefixo, SF – sufixo, IV – invariante, IF – Infixo, PPF – Prefixo próprio, SSF – Sufixo próprio;
(5) "<l>" e "</l>" que determina o idioma do lexema, sendo 1 para o alemão, 2 para o inglês, 3 para o português, 4 para o espanhol, 5 para o francês e 6 para o sueco.

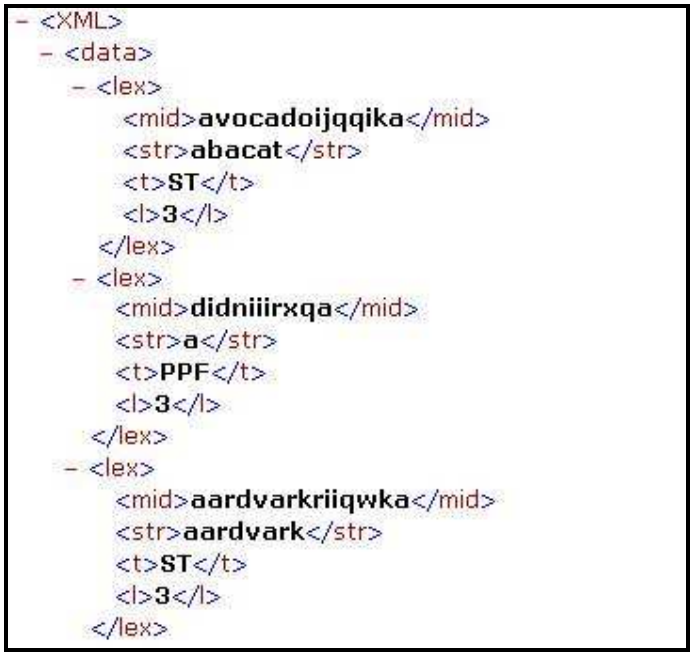



Figura 8: Representação da estrutura XML de três lexemas do tesauro *Morphosaurus*

## 2.10 O SISTEMA *MORPHOSAURUS*

É corriqueiro assumir que a forma de comunicação idiomática entre seres humanos está centrada na palavra. Isto se baseia na hipótese de que o arranjo de palavras ou termos são unidades básicas para a construção de frases e sentenças (SCHULZ, 2006). Na Teoria Sintática, a palavra é o símbolo final para a representação da linguagem e, num primeiro momento, poder-se-ia afirmar que a palavra seria a representação final dos objetos do mundo real através da linguagem natural. Entretanto, verificando-se o sentido das expressões da linguagem natural, especialmente em linguagens técnicas, pode-se encontrar evidências de que a atomicidade semântica freqüentemente não coincide ao nível da palavra em si. Por exemplo, a palavra em português "embaraçada": esse termo traduzido para a língua espanhola não traduz o mesmo sentido. Em línguas técnicas como a médica, os sentidos atômicos são encontrados em diferentes níveis de fragmentação e granularidade. Um sentido atômico pode corresponder a um radical (e.g., "hepat" referente ao "fígado"), prefixo (e.g., "anti-", "hipo-", "des"), sufixo (e.g., "ose", "ite", "logia"), fragmentos de palavras longas (e.g., "neurosis", "hipofis-"), a própria palavra (e.g., "pé", "pais"). Termos compostos – tanto multi-palavra (e.g., "vitamina C") quanto aglutinadas (e.g.; "hipofis-"), freqüentemente têm um sentido próprio que não pode ser deduzido do significado dos componentes. Ainda pior, o sentido do

termo composicional pode, às vezes, contradizer o sentido literal dos componentes: por exemplo, a doença “mycosis fungoides” não é uma micose, e “neurose” não é uma doença dos nervos. As possibilidades de combinações para a formação de palavras é extensa e a formação de termos é comum. Como conseqüência, uma boa cobertura de um léxico a um determinado domínio ocorreria somente se suas unidades lexicais pudessem restritar a um sentido atômico, as quais podem ser utilizadas como blocos para a formação de termos em qualquer nível de granularidade. Desta forma, possibilita-se extrair sentido atômico de textos no sentido de formar uma base para interpretação semântica de textos em linguagem natural, muito importante para aplicações no campo da Recuperação de Documentos, Extração de Informação e Mineração de Texto.

As coleções de documentos na área de saúde são imensas e dinâmicas. Do ponto de vista da RI, isso dificulta a reutilização de muitas abordagens que desempenham eficientemente sob condições experimentais de pequena escala como a indexação semântica latente ou modelos probabilísticos ainda mais sofisticados (FUHR, 1992). Isso ocorre porque nenhuma máquina de busca ainda é capaz de manter vetores de índices de documentos de grandes dimensões ($n >> 100.000$) para grandes volumes de documentos e altas freqüências de atualização (HAHN, SCHULZ, MARKÓ *et al.*, 2004).

Além disso, as coleções de documentos são multilíngües. Apesar dos documentos clínicos serem tipicamente escritas nas referidas línguas nativas, as buscas nas maiores bases bibliográficas, como o MEDLINE, requer conhecimento da terminologia médica na língua inglesa que só parte dos profissionais da saúde detêm. Portanto, é necessário um mecanismo que realize uma ponte entre a comunicação tanto a nível de sistema como de usuário entre as diversas bases de conhecimento em saúde distribuído.

A população de usuários de sistemas de recuperação de documentos médicos é heterogênea, mesmo tratando-se de uma mesma especialidade. Logo, a implantação de um sistema que realize o mapeamento entre as diferentes terminologias parece ser uma idéia inevitável para satisfazer às necessidades das diversas comunidades como estratégia de busca heterogênea. Por isso, a simplicidade na representação dos documentos, bem como o seu mapeamento conceitual e, num nível mais detalhado, o mapeamento lexical de forma intra e interlingual tornam-se questões cruciais para uma metodologia que se propõe ser adequada para um sistema de recuperação de documentos médicos, e inclusive, para um Sistema de Recuperação de Informações Multilíngüe.

Na base desses dois desafios para RI, o multiligüismo de um lado, e a granularidade semântica de outro lado, surgiu o Sistema *Morphosaurus*, com o intuito de responder com

uma metodologia em que são empregados descritores artificiais representativos dos conceitos da terminologia médica baseada num tesauro multilíngüe que consolida uma abordagem em que são empregadas unidades lexicais semanticamente atômicas. A indexação de documentos baseados em entidades lexicais semanticamente atômicas evita a explosão de um léxico quando se procura contemplar todas as variantes morfológicas de uma palavra (SCHULZ e HAHN, 2000).

### 2.10.1 Tesauro de *Subwords*

O Sistema *Morphosaurus* é uma ferramenta de recuperação de documentos médicos desenvolvido pelo Departamento de Informática Médica da Universidade de Freiburg em cooperação com o Laboratório de Engenharia de Informação e Línguas da Universidade de Jena e o Programa de Pós-Graduação em Tecnologia em Saúde da Universidade Católica do Paraná.

A maior particularidade do tesauro utilizado é que suas entradas lexicais correspondem, em grande parte, ao que foi definido como *subword*. *Subwords* não são termos e na sua maioria, não são palavras que possam ser encontradas em textos. Na sua maioria, *subwords* correspondem a morfemas ou grupos de morfemas. O critério fundamental para a delimitação de *subwords* é que representem conceitos atômicos relevantes do domínio.

### 2.10.2 Atomicidade Semântica

Na Teoria Lingüística, uma seqüência de caracteres é considerada semanticamente atômica se o seu significado, seu sentido, não deriva unicamente de seus morfemas constituintes, seja por inflexão, derivação ou composição na formação de uma palavra; ou seja, ela por si só é representativa de um significado. As palavras são formadas através de processos morfológicos como inflexão, derivação e composição. Por exemplo, “neurose” é o resultado da ligação de “nerv” (nervo) com “ose” (doença). Entretanto, o sentido de “neurose” não significa doença de nervo(s). Conseqüentemente, pode-se considerar a derivação “neuros” como uma unidade lexical atômica a ser acrescida no léxico do tesauro português.

Unidades lexicais podem ter múltiplos sentidos (homonímia) e o sentido pode ser expresso de formas diferentes (sinonímia). Apesar de se construir terminologias específicas para determinados domínios com objetivo de controlar o uso de linguagem especializada e evitar expressões ambíguas, terminologias não padronizadas, em vários domínios, ainda são

utilizadas. Por exemplo, o radical inglês "*head*" possui diferentes sentidos: "*headache*", "*head of femur*" ou "*head of departament*". O mesmo pode ser notado para a palavra "operação" que pode significar um "procedimento cirúrgico" no domínio médico em oposição aos outros sentidos como no domínio da matemática ou dos negócios. Nos casos citados, o contexto geralmente ajuda a selecionar o verdadeiro sentido. Além disso, um domínio muito bem delimitado ("*restriction to a well-defined domain*") permite ignorar outros sentidos definitivamente pertencentes a outros domínios (isto é, "head" como parte de uma frase na teoria gramatical).

Além de ambigüidades, algumas unidades lexicais possuem sobreposição de sentido (*overlapping senses*). Relações quase-sinônimas podem ser verificadas entre termos de línguas diferentes (Latim "*caput*" vs. Inglês "*head*") ou diferentes níveis de erudição ("*belly*" vs. "abdomen"). Raros são os casos de equivalência total, isto é, sinonímia perfeita em todos os contextos.

Para se estabelecer classes de sinônimos, primeiramente deve haver um compromisso claro a respeito das expressões que podem ser consideradas sinônimos; isto manterá a integridade do contexto do domínio. Em segundo lugar, com relação às delimitações semânticas, deve-se manter a compatibilidade com as propriedades formais das relações de equivalências, isto é, manter as propriedades de reflexividade, transitividade e simetria: se um lexicógrafo considera que "*disease*" é sinônimo de "*illness*" e este sinônimo de "*sickness*", então "*disease*" e "*sickness*" são sinônimos. Delimitações semânticas irão depender em resolver relevância de sentidos distintos e sutis no contexto do domínio considerado. Por exemplo, para um leigo, as palavras do domínio da medicina "neoplasma", "carcinoma" e "câncer" podem ser considerados como sinônimos, mas não para um profissional de saúde. Outro exemplo seria equalizar "estirp-", "remov-", "ectom-" de forma geral no domínio médico negligenciando distinções sutis das técnicas cirúrgicas.

Como dito e escrito, o tesauro do Sistema *Morphosaurus* é multilíngüe. Isso envolve tradução, como caso especial de sinonímia, no qual as palavras em diferentes línguas são ligadas através das relações de equivalências. Assim, os termos, por exemplo, em inglês, "*disease*" e "*illnees*", com os termos alemães, "*krankheit*", espanhol, "*enfermedad*", francês, "*maladie*", sueco, "*sjukdom*" e português, "doença" são todos reunidos na mesma classe de equivalência para representar o mesmo sentido.

A delimitação de uma classe de sinônimo depende do contexto do domínio, por exemplo, "leucemia" e "neurose" significam literalmente "sangue branco" e "doença do nervo". Esses termos possuem origens históricas e que, atualmente, não provê descrição

completa quando relacionada à medicina moderna. Assim, tem-se um termo, originalmente composto, formado por um contexto histórico, mas que, atualmente, é tomado como um sentido atômico.

No sistema *Morphosaurus*, para representar os sentidos atômicos das unidades lexicais, as classes de equivalências são mapeadas para uma camada de descritores independentes do idioma, nomeadamente, **ID**entificadores *Morphosaurus* (***M****orphosaurus* ***ID****entifiers* - **MID**s). Os símbolos dessa representação referem-se a todas unidades lexicais que possuem o mesmo sentido nas línguas consideradas. Classes de equivalências podem ser grosseiramente consideradas como conceitos em tesauros, como as *synsets* no *Wordnet* (FELLBAUM, 1998) ou, no domínio médico, como as "***c****oncept* ***u****nique* ***i****dentifier* – **CUI**s" do *Metathesaurus* da *Unified Medical Language System* – UMLS (UMLS, 2005), a qual atualmente, através de um sistema hierárquico interligado combina mais que uma centena de terminologias médicas heterogêneas (tesauros, classificações, etc), entre outros, a Classificação Internacional de Doenças – CID, o "*Medical Subject Heading* – MeSH", o SNOMED CT, etc...

Entretanto, existem duas diferenças básicas entre as MIDs e os *Synsets* e CUIs: primeiramente, as MIDs também podem representar disjunções de diferentes sentidos; é o caso quando se depara com unidades lexicais ambíguas. Como exemplo, pode-se citar o caso do termo "molar". Esse termo é ambíguo, assim como muitos sinônimos e traduções do mesmo, e é representado por uma MID. Essa MID (ambígua) é relacionada com pelo menos duas MIDs as quais representam os sentidos, pela relação "*has_sense*". Em segundo lugar, a MID é um descritor único e, assim sendo, é possível realizar qualquer tipo de arranjo com outras MIDS de forma a assumir qualquer tipo de configuração de relacionamento sintagmático ou paradigmático.

### 2.10.3 Indexação Morfossemântica

Uma *subword* é uma unidade lexical mínima significativa de um termo de um certo domínio. Essa premissa define a propriedade de que o sentido não pode ser decomposto. Desta forma, pode-se considerar o termo hepat⊕ite[6] como composição de duas *subwords*, "hapat" e "ite", pois o seu sentido, a sua interpretação é decorrente de seus constituintes; em oposição à, por exemplo, hipo⊕fise, que é semanticamente indeterminada, pois a

[6] O símbolo ⊕ é utilizado para separar termos semanticamente atômicos que compõem uma palavra.

interpretação de seus constituintes individualmente não representa o sentido verdadeiro de hipófise.

Uma *subword* pode ser um *stem* (ST), um prefixo (PF), um sufixo (SF), ou uma invariante (IV). Ainda nesse sistema, definiu-se o prefixo e sufixo próprio, os quais não podem ser prefixados ou sufixados:

(1) Os *stems* (ST) como "gastr", "hepat", "*diaphys*", "neuros" são considerados como a parte principal de uma palavra com maior peso semântico. Os *stems* podem ou não serem prefixados, sufixados, ou ainda ocorrerem sem afixo;

(2) Prefixos (PF) como "de", "re", "in", "*hyper*", "anti-", precedem aos *stems* uma ou mais vezes[7];

(3) Prefixos Próprios (PP) como "peri-", "hemi-", "*down*-" são prefixos que não podem ser pré-fixados por outros prefixo;

(4) Infixos (IF) como "gastr-o-intestinal", ou "- r -" em "hernio-r-rafia" são usados como um ente de ligação entres alguns *stems* na formação de palavras compostas;

(5) Sufixo (SF) como "-a", "-io", "-ion", "-tomia", "-ite" complementam (seguem) um *stem* ou um outro sufixo;

(6) Sufixos Próprios (SP) como "-ação", "-ão", "-essemos", são sufixos que não podem ser seguidos por outros sufixos.

Não só *stems*, mas também muitos prefixos e sufixos como "anti-", "-logia" e "-itis", têm uma relevância para a indexação e não podem ser ignorados. Outros, como terminações de plural ou de tempos verbais podem ser ignoradas. Os lexemas desse tipo, classificados como sufixos, são utilizados para segmentar os casos.

Há casos que justificam a introdução de uma nova categoria: (1) lexemas que sempre correspondem a palavras inteiras e não podem ser flexionados, derivados ou compostos como é o de acrônimos como "ECG" ou "AVE"; (2) strings muito curtas como "ion" ou "gen", que devido a sua ocorrência como partes de muitas palavras produzem muitas segmentações errôneas. Em ambos os casos, as entidades lexicais são classificadas como invariantes (IV).

O léxico e tesauro do Sistema *Morphosaurus* é utilizado como uma base de conhecimento semântico para a geração de uma linguagem artificial, composta por MIDs; e, para isso, utiliza-se de um modelo de *subword* que pode ser representada por um autômato de estado finito, conforme ilustra a figura 8. Esse modelo expressa que uma palavra pode opcionalmente ser segmentada começando por prefixo, seguido de um *stem* (que pode ser

[7] Em geral, para o termo "hemi ⊕ an ⊕opsia", o prefixo "an" é prefixado pelo termo "hemi".

seguido por outro *stem*, separado por um infixo e/ou adicionalmente seguido por um prefixo ou sufixo) e terminado por sufixo (próprio).

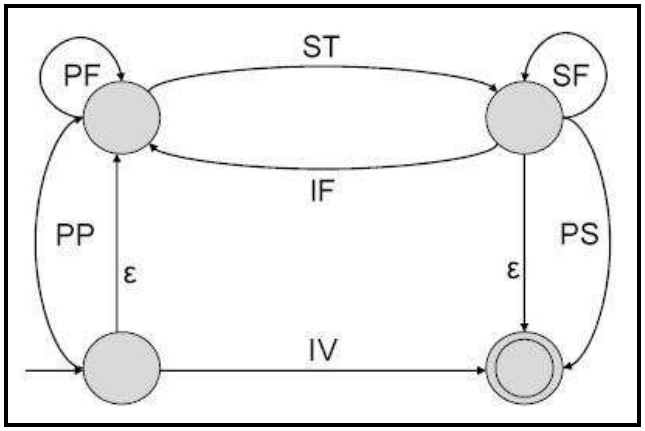



Figura 9: Autômato de estados-finitos para o modelo de *Subword* do Sistema MS. ST: stem, PF: prefixo, PP: prefixo próprio, IF: infixo, SF: sufixo, SP: sufixo próprio.

2.10.3.1 Caracterização do Léxico de Subwords

Pode-se formalizar as características de uma *subword* da seguinte forma: define-se ***LU***:={gastr, hepat, figad, ...} como itens lexicais a nível de *subwords*. Fixa-se ***M*** := {#gastr, *#liver*, #inflam, ...} como um conjunto de classes de equivalência, MID e convenciona-se a anotação de cada classe de equivalência com o símbolo "#" seguido de uma entrada lexical não ambígua, válida. Considera-se ***T*** := {PP, PF, ST, IV, SF, SP} que denotando o tipo da *subword* como descrito anteriormente. Define-se, também, ***L*** := {EN, GE, FR, SP, PT, SW} para representar os idiomas contemplados no tesauro (Inglês, Alemão, Francês, Espanhol, Português e Sueco, respectivamente) e ***D*** := {Medicina Clínica, Indústria Farmacológica, Odontologia, Veterinária, etc...} para representar o contexto do domínio. Então, a relação entre unidade lexical, sentido, tipo de lexema, domínio do contexto e idioma como sendo expresso por um quíntuplo (***LU, M, T, D, L***).

Se nenhum significado for assinalado para a entrada lexical, então, esta é considerada como uma "*stop entry*" (termo não considerado no processamento); tendo somente uma função gramatical, como, por exemplo, os verbos auxiliares e terminações utilizadas nas inflexões das palavras. A seguir, apresentam-se alguns exemplos típicos de entradas lexicais com seus atributos e relacionamentos:

(1) Sinônimos: o sufixo inglês "-itic" e "-itis" possuem o mesmo sentido de "*inflammation*":

i. $l_1$ = (*inflamm*, ST, #*inflamm*, EN, $d_1$)
ii. $l_2$ = (*itic*, SF, #*inflamm*, EN, $d_1$)
iii. $l_3$ = (*itis*, SF, #*inflamm*, EN, $d_1$).

(2) Tradução: o *stem* alemão *entzünd* (transcrito para *entzuend*) e o sufixo em português "-ite" denota o mesmo sentido do *stem* inglês "*inflamm*":

i. $l_1$ =(*inflamm*, ST, #*inflamm*, EN, $d_1$)
ii. $l_4$ =(*entzuend*, ST, #*inflamm*, GE, $d_1$)
iii. $l_5$ =(*ite*, SF, #*inflamm*, PT, $d_1$).

(3) Ambigüidade: o substantivo "*head*" em inglês, ou "cabeça" em português, pode se referir a uma parte anatômica do corpo como a uma pessoa:

i. $L_6$ =(*head*, ST, #$head_1$, EN, $d_1$)
ii. $L_7$ =(*head*, ST, #$head_2$, EN, $d_1$).

(4) *Stop Entries*: a palavra "*era*" é um substantivo em inglês, mas um verbo auxiliar nas línguas latinas como espanhol e português:

i. $L_8$ = (*era*, ST, #*era*, EN, $d_1$)
ii. $L_9$ = (*era*, IV, ε , SP, $d_1$)
iii. $L_{10}$ = (*era*, IV, ε , PT, $d_1$).

(5) Quase-sinônimos: as palavras "sildenafil" e o nome "viagra" podem ser considerados sinônimos no campo da medicina clínica ($d_1$), mas não no campo da indústria farmacêutica ($d_2$):

i. $L_{11}$ = (*sildenafil*, ST, #*sildenafil* , EN, $d_1$)
ii. $L_{12}$ = (*viagra*, IV, #*sildenafil* , EN, $d_1$)

iii. $L_{13}$ = (*sildenafil*, ST, #*sildenafil* , EN, $d_2$)

iv. $L_{14}$ = (*viagra*, IV, #*viagra*, EN, $d_2$).

2.10.3.2 Tesauro de *Subwords*

O tesauro organiza as classes de equivalências tanto de forma monolíngüe quanto de forma multilíngüe. As entradas lexicais que compartilham da mesma MID pertencem à mesma classe de equivalência, ou seja, a classe de equivalência é um subconjunto das entradas lexicais: $\boldsymbol{C} \subset \boldsymbol{E}$ . Por convenção, anota-se esse conjunto com *c* seguido pelo seu símbolo correspondente (MID) na forma subscrita. Por exemplo, o conjunto $c_{\#\text{inflamm}}$ contém todos os itens lexicais das diferentes línguas que possuem o sentido de inflamação:

$c_{\#\text{inflamm}}$ := { (*inflamm*, ST, #*inflamm*, EN, $d_1$),
(*itic*, SF, #*inflamm*, EN, $d_1$),
(*itis*, SF, #*inflamm*, EN, $d_1$),
(*entzuend*, ST, #*inflamm*, GE, $d_1$),
(*ite*, SF, #*inflamm*, PT, $d_1$),
…
}.

Tesauros podem suportar vários tipos de relacionamentos. Porém, neste trabalho foram contemplados basicamente dois tipos: a relação do tipo horizontal "*has_word_part*" $\subset \boldsymbol{M}$ x $\boldsymbol{M}$ (relação sintagmática) e a relação do tipo vertical "*has_sense*" $\subset \boldsymbol{M}$ x $\boldsymbol{M}$ (relação paradigmática).

Pode-se definir $\boldsymbol{R}$ como um conjunto de MIDs relacionadas. Então:

(1) O conjunto $\boldsymbol{R_1} := \{(m_o, m_1), (m_o, m_2), (m_o, m_3), ..., (m_o, m_n),\} \in$ "*has_word_part*" (com $m_{1, \ldots, n} \in \boldsymbol{M}$ e $|\boldsymbol{R_1}| \geq 2$) relaciona um MID $m_o$ a uma lista de pelo menos duas outras MIDs. Este tipo de relacionamento sintagmático é utilizado no sentido de "esconder" as composições semânticas explícitas; e.g., expandir aos termos atômicos que compõem um acrônimo, por exemplo; ou ainda, para indicar os termos componentes quando há supressão de caracteres na composição de uma palavra, como por exemplo "urinálise":

i. $\boldsymbol{R_1}$ := {(*#urinalis,#urin*), (*#urinalis, #analis*) } ∈ "*has_word_part*"

(2) O conjunto $\boldsymbol{R_2}$ **:=** {($m_o$, $m_1$), ($m_o$, $m_2$), ($m_o$, $m_3$), ..., ($m_o$, $m_n$),} ∈ "*has_sense*" (com $m_{1, \ldots,n} \in \boldsymbol{M}$ e | $\boldsymbol{R_1}$| ≥ 2) relaciona um MID $m_o$ ambíguo a um conjunto de pelo menos duas MIDs. Esse tipo de relacionamento é utilizado para indicar as possíveis acepções (não ambíguas) do MID ambíguo. Como exemplo, pode-se citar o MID ambíguo $c_{\#cabec}$ que podemos representar por:

i. $\boldsymbol{R_1}$ := {(*#cabec,#head*), (*#cabec, #chief*) } ∈ "has_sense"

(3) Juntos, ambos relacionamentos constituem o thesaurus τ de um domínio ***D***.

$$\tau_{\boldsymbol{D}} := (\textit{has\_word\_part}, \textit{has\_sense})$$

2.10.3.3 Indexação das *Subwords*

O tesauro de *subwords* é um recurso declarado para a normalização morfossemântica de textos no domínio da medicina. O terceiro componente, o indexador, diz respeito aos procedimentos relacionados à normalização e à indexação como módulos do Sistema *Morphosaurus*. As palavras de um texto de uma dada língua são transcritas para uma linguagem artificial representadas pelas MIDs. Esse procedimento segue três etapas seqüenciais. A tabela 2 ilustra exemplos baseados no mesmo texto para português, alemão e inglês. Esses procedimentos foram realizados por rotinas desenvolvidas nas linguagens de programação PERL e JAVA, implementadas na plataforma *Unix/Linux* disparados por alguns *scripts* em "*Shell Bash*". Além disso, o tesauro é utilizado no padrão XML convertido do léxico na base de dados MySQL.

I. Normalização Ortográfica

A tabela 2 indica como os documentos são convertidos em representações multilíngüe realizadas em três etapas. O primeiro passo trata da normalização ortográfica. Um pré-processador converte todos os caracteres capitalizados para minúsculo e realiza substituições de caracteres específicos para cada língua, de forma a facilitar a equivalência entre os *tokens* de texto e as entradas do léxico. Por exemplo, no alemão troca-se: "β" → "ss", "ä" → "ae",

“ö” → “oe”, “ü” → “ue” e, no português, “ç” → “c”, “ú” → “u”, “õ” → “o”, e assim por diante. Ainda no alemão há um procedimento adicional motivado pela idiossincrasia da linguagem médica, e.g., no caso da língua alemã: “ca” → “ka”, “co” → “ko”, “cu” → “ku”, “ce” → “ze”, “ci” → “zi”, etc. Isso resolve um problema notório na terminologia médica ale mã, em que alguns termos originais em latim que contêm a letra “c” ao invés de “k” e “z”, não permitem o uso do “c” para os mesmos termos derivados em alemão. Essa regra é freqüentemente quebrada até mesmo pelos profissionais da saúde (ou seja, em alemão são utilizadas formas gráficas diferentes para o mesmo sentido: “*karzinom*”, “*carzinom*”, “*carcinom*”).

II. Segmentação Morfológica

Após a normalização ortográfica, o sistema decompõe o texto normalizado ortograficamente em uma seqüência de *subwords* (correspondentes às entradas no léxico) e restos lexicais (não presente no léxico) (ANDRADE, NOGUEIRA-NETO, SCHULZ *et al.*, 2004). O resultado da segmentação é verificado por um autômato finito que rejeita segmentações inválidas (isto é, sem lemas ou que se iniciam com sufixo). Se existirem leituras válidas ambíguas ou segmentações incompletas devido a entradas inexistentes no léxico, regras são aplicadas para encontrar as segmentações mais longas, com o menor número de segmentos não especificados, etc. Se o algoritmo de segmentação não detectar uma leitura válida, a palavra original é restituída.

III. Normalização Semântica

Nesse passo final, cada *subword* é substituída pelo seu MID. Depois disso, todos os sinônimos de uma mesma língua e todas as traduções de *subwords* que se equivalem semanticamente em línguas diferentes são representadas pelo mesmo item de código na representação final. Os termos compostos (como ‘myalg⊕y’), que são ligados aos seus componentes por meio da relação ‘*has_word_part*’, são substituídos pelas MIDs dos seus componentes. As classes ambíguas, aquelas relacionadas por “*has_sense*”, resultam numa seqüência de MIDs de termos ambíguos. O resultado é um documento normalizado morfossemanticamente em uma representação multilíngüe, independente de língua (ANDRADE, NOGUEIRA-NETO, SCHULZ *et al.*, 2004).

Tabela 2: Normalização Morfossemântica para o mesmo texto em inglês, alemão e português.

| **Documento Original** | **Normalização Ortográfica** | **Segmentação Morfológica** | **Normalização Semântica** |
|---|---|---|---|
| High TSH values suggest the diagnosis of primary Hypothyroidism | high tsh values suggest the diagnosis of primary hypothyroidism | high tsh value s suggest the diagnos is of primar y hypo thyroid ism | **top# tsh# value#** suggest# **diagnos# first# hypo# thyroid#** |
| Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose | erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose | Er hoeh te tsh – wert e erlaub en die diagnos e einer primaer en hypo thyre ose | **top# tsh# value#** allow# **diagnos# first# hypo# thyroid#** |
| A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário | a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario | a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primary o | current# **value# top# tsh#** suggest# **diagnos# hypo# thyroid# first#** |

### 2.10.4 Implementação do Modelo de *Subwords*

O modelo de *subwords* implementado dentro do sistema *Morphosaurus* contempla a terminologia no domínio da medicina clínica nas seguintes línguas inglesa, portuguesa, alemã, sueca, espanhola e francesa. A seguir, segue a explicação da estratégia para a criação, manutenção e validação do léxico e do tesauro.

#### 2.10.4.1 Criação do Léxico

O início da construção do léxico de *subwords* ocorreu por meio de uma lista padronizada de afixos compilados do domínio em questão. Elas puderam ser obtidas de especificações morfológicas da gramática de cada língua que serviu de base para estabelecer a delimitação do *stem* de uma palavra pela compatibilidade com o próprio prefixo ou sufixo. Com a aplicação de técnicas de estado-finito em grandes quantidades de textos, é possível gerar, das palavras, grandes quantidades de termos candidatas à *subword* pela decomposição das palavras. Obviamente, após o processo de inclusão, torna-se necessária a validação da mesma de forma empírica, pelos lexicógrafos.

- Delimitação da *Subword*

O processo de delimitação, em termos de seqüência de *string* para definir uma *subword* está embutido de interpretação do resultado que termo possa carregar. A motivação para a delimitação de uma *subword* é função do resultado da própria segmentação. Por exemplo, para a *subword* do léxico, no domínio D, "nefrotomia" pode ser segmentada na forma:

(*nefr*, ST, #*kidney*, EN, $d_1$) ⊕ (*o*, IN, #ε, EN, $d_1$) ⊕ (*tomy*, SP, #*incision*, EN, $d_1$).

Mas também poderia ser:

(*nefr*, ST, #*kidney*, EN, $d_1$) ⊕ (*oto*, ST, #*ear*, EN, $d_1$) ⊕ (*my*, ST, #*muscle*, EN, $d_1$).

As rotinas de segmentação formam configuradas para escolher, entre outros critérios, primeiramente os termos mais longos. Às vezes, o resultado não expressa o verdadeiro significado; então, uma solução pragmática foi incluir um outro lexema sinônimo, de forma a manter a integridade do significado, como, por exemplo: $e_{14}$ = (*nephro*, ST, #*kidney*, EN, $d_1$); e suas variantes em outras línguas: $e_{15}$ = (*nefro*, ST, #*kidney*, PT, $d_1$), $e_{16}$ = (*nefro*, ST, #*kidney*, SP, $d_1$).

- Validação empírica *de string* específica

A validação de *subwords*, especialmente de *stems* curtos como é o caso de "gen", "ship", "mi", é propenso a efeitos "colaterais" como descrito anteriormente. Pelo fato de serem muito curtos, é freqüente a produção de *substrings* de forma a gerar uma segmentação errada. Para verificar possíveis erros de segmentações causados por esses tipos de *subwords*; listas compiladas de textos são submetidas ao sistema para serem verificadas e corrigidas pelos lexicógrafos. Dois casos são verificados: problemas com palavras curtas e segmentações erradas; e critérios para inclusão de *subwords* no léxico.

A seleção de unidades lexicais deve refletir a domínio em questão. Para isso, listas de palavras selecionadas estatisticamente de *corpora* foram utilizadas para medir a relevância dos termos. O ideal é que cada entrada lexical corresponda a uma entidade indivisivelmente

atômica correspondente a uma unidade semântica. Entretanto, há exceções, especialmente para os casos de lexemas compostos que possuem um sentido atômico. Como conseqüência pode-se citar que (1) uma entrada lexical pode ser uma palavra inteira, (2) um termo que pode ser "expandido" aos seus significados atômicos pela relação sintagmática "*has_word_part*". Por exemplo, o termo "Ascórbico" implica os termos "C" e "vitamina". Um caso especial para isso são os acrônimos (AVE, ECG, etc...).

Situação 1:

$e_{22}$ = (*ascorb*, ST, #*ascorb*, EN, *d*)
$e_{23}$ = (*vitamin c*, IV, #*ascorb*, EN, *d*), e

Situação 2:

$\boldsymbol{R_1}$ := {(#*ECG,#electro*), (#*ECG, #cardi*), (#*ECG, #gram*) } ∈ "*has_word_part*"

Por razões de parcimônia, termos compostos geralmente não são incluídos no léxico, a não ser que esses termos tenham um sinônimo não divisível (e.g., *vitamin C*) ou que o significado das palavras contradiz o significado do termo (e.g., *mycosis fungoides*). Em muitos outros casos o significado de termos compostos não é exatamente decorrente do significado de seus componentes, mas seus componentes são literalmente traduzidos para as outras línguas. Nesses casos, a inclusão de termos compostos não é admitida (e.g., febre amarela).

Os nomes próprios são incluídos no léxico sob as seguintes circunstâncias:

(1) relacionando os sinônimos entre os diferentes nomes de produtos, por exemplo:

$e_{24}$ = (*diclofenac*, S, #*diclofenac*, EN, *d*)
$e_{25}$ = (*voltaren*, S, #*diclofenac*, EN, *d*)
$e_{26}$ = (*cataflam*, S,, #*diclofenac*, EN, *d*);

(2) quando são utilizados como epônimos, isto é, pertencem ao mesmo domínio terminológico:

$e_{27}$ = (*crohn*, S,, #*crohn*, EN, *d*)

$e_{28}$ = (*parkinson*, S, #*parkinson*, EN, *d*) [8];

(3) quando existe a tradução, especialmente de termos geográficos:

$e_{29}$ = (*switzerland*, I, #*switzerland*, EN, *d*)

$e_{30}$ = (*suisse*, S, #*switzerland*, FR, *d*)

$e_{31}$ = (*suiç*, S, #*switzerland*, BR, *d*).

2.10.4.2 Criação do Tesauro

Uma classe de equivalência reúne as variações morfológicas de um lexema para estabelecer a definição de um mesmo sentido tanto de forma monolíngüe quanto de forma multilíngüe. Para essa classe de equivalência, estabelece-se um único MID.

A criação do tesauro acontece por meio de dois tipos de relações para estabelecer vínculos entre as classes de equivalências. Uma relação sintagmática, pela relação "*has_word_part*" e pela relação paradigmática, pela relação "*has_sense*. A figura 10 apresenta um exemplo abordando as duas situações.

A figura 10 contempla um exemplo para dois casos: a relação paradigmática liga um MID ambíguo a outros sentidos; e.g., *#head* é ligado aos MIDs *#caput={"cabec-", "kopf", ...}* e *#boss = {"chief", "haeupt"}* pela relação *has_sense*; enquanto que a relação sintagmática "*has_word_part*" realiza a ligação de um MID aos seus MIDS "atomicamente" semânticos; e.g., o MID *#myalg* ={"*myal-*", "mialg-", ...} aos MIDs *#muscle* = {"*myo-*", "mio-", "muscul-", ...} e *#pain* = {*"pain", "algy-"*, "dor", *"schmerz", ...*}. A razão desse tipo de relacionamento é evitar uma segmentação errônea pelo fato de se tratar do *stem "myo"* muito curto.

O relacionamento tipo "*has_sense*" relaciona as possíveis acepções de uma classe ambígua enquanto o relacionamento do tipo "*has_word_part*" conecta uma MID a outras possíveis MIDs atômicas distintas que fazem parte da interpretação conjunta para um mesmo sentido. Esse procedimento é realizado nas seguintes situações:

1. morfemas muito curtos: devido à sua heurística, a segmentação pode levar a outras interpretações errôneas. Por exemplo:
   $e_{24}$ = (*myalg*, ST, #*myalg*, EN, *d*)

[8] Alguns nomes próprios podem ser morfologicamente alterados, e.g. "parkinsoniano".

$e_{25}$ = (*mialg*, ST, *#myalg*, PT, *d*)

$e_{26}$ = (*muscl*, ST, *#muscle*, EN, *d*)

$e_{26}$ = (*muscl*, ST, *#muscle*, PT, *d*)

$e_{26}$ = (*pain*, ST, *#pain*, EN, *d*)

$e_{26}$ = (*algia*, ST, *#pain*, PT, *d*),

ou resumindo:

$\boldsymbol{R_3}$ := {(*#myalg,#muscle*), (*#myalg, #pain*) } ∈ "*has_word_part*";

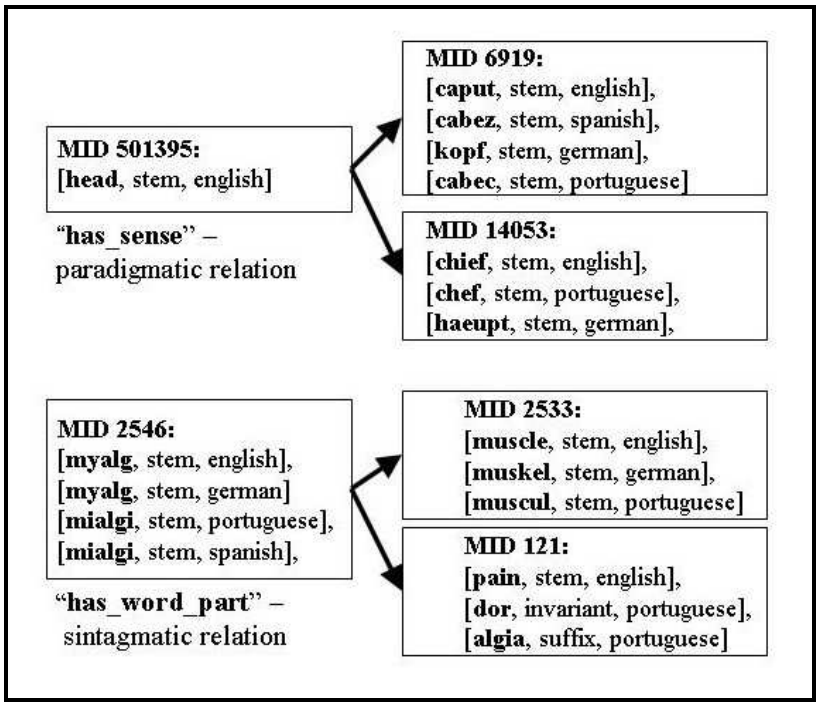



Figura 10: Tipos de Relacionamento semânticos suportados pelo tesauro do MS.

2. um lexema que é atômico em um idioma, mas composicional em outro:
   $\boldsymbol{R_4}$ := {(*#esparadrap,#adhesiv*), (*#esparadrap, #tape*) } ∈ "*has_word_part*";

3. quando há uma contração na formação de uma palavra, por exemplo, no português, para o termo contraído "urinálise" ocorre a perda da letra "a". Então, a solução para boa segmentação é relacioná-lo na forma abaixo:
   $\boldsymbol{R_5}$ := {(*#urinalis,#urina*), (*#urinalis, #analis*) } ∈ "*has_word_part*".

2.10.4.3 Aspectos da Criação do Tesauro

A delimitação de classes semânticas é uma tarefa totalmente intelectual que provavelmente não pode ser automatizada (SCHULZ, 2006) e pressupõe um excelente conhecimento da terminologia do domínio. Entretanto, como ponto de partida, cada lexema possui uma única MID. Se um lexicógrafo concluir que duas ou mais entradas possuem o mesmo sentido, então elas são unidas como sinônimos.

O procedimento de juntar ou não classes de equivalências talvez seja o trabalho mais complicado para um lexicógrafo. Observa-se, por exemplo, o caso de "tumor", "sarcoma", "câncer" e "carcinoma". Como contemplá-los num tesauro? Um leigo poderia considerá-los como sinônimos, mas não um profissional de saúde, para os quais os sentidos desses termos podem parcialmente se sobrepor.

2.10.5 Editor do Tesauro *Morphosaurus – MorphoEditWeb*

O *MorphoEditWeb* é a ferramenta principal para os trabalhos de edição e manutenção do tesauro que foi desenvolvida como uma solução cliente-servidor de tal forma que diversos usuários de diferentes lugares possam acessar e manipular o mesmo repositório de *subwords* através da Internet [9]. A figura 11 apresenta a interface *MorphoEditWeb*.

Inicialmente, a ferramenta foi desenvolvida em Visual Basic e usada em um ambiente Windows multi-usuário (Servidor Citrix). Com o crescimento do grupo de lexicógrafos, sentiu-se a necessidade de torná-la disponível via Internet. Desta forma, o *MorphoEditWeb* foi refeito na linguagem JAVA, utilizando o MySQL como banco de dados. O processamento é feito em cima do léxico no padrão XML e cada lexicógrafo possui sua própria conta.

2.10.5.1 Fontes de Terminologias como Ferramentas de Apoio

Além das funcionalidades inerentes ao gerenciamento do léxico tesauro, foram incluídas algumas ferramentas de apoio à decisão dos lexicógrafos. Essas ferramentas de apoio, compiladas de bases de terminologia mundialmente reconhecidas, oferecem informações de forma a servir de subsídio no sentido de decisão em delimitar uma *subword*, sobre a relevância do lexema, relacionar um outro significado, sinônimo e assim por diante.

[9] http://morphwww.medinf.uni-freiburg.de:8080/MEWeb.

Elas mostram os significados de forma multilingüe além de apresentar outras formas de escrever o mesmo conceito. As informações foram compiladas do UMLS e do MESH.

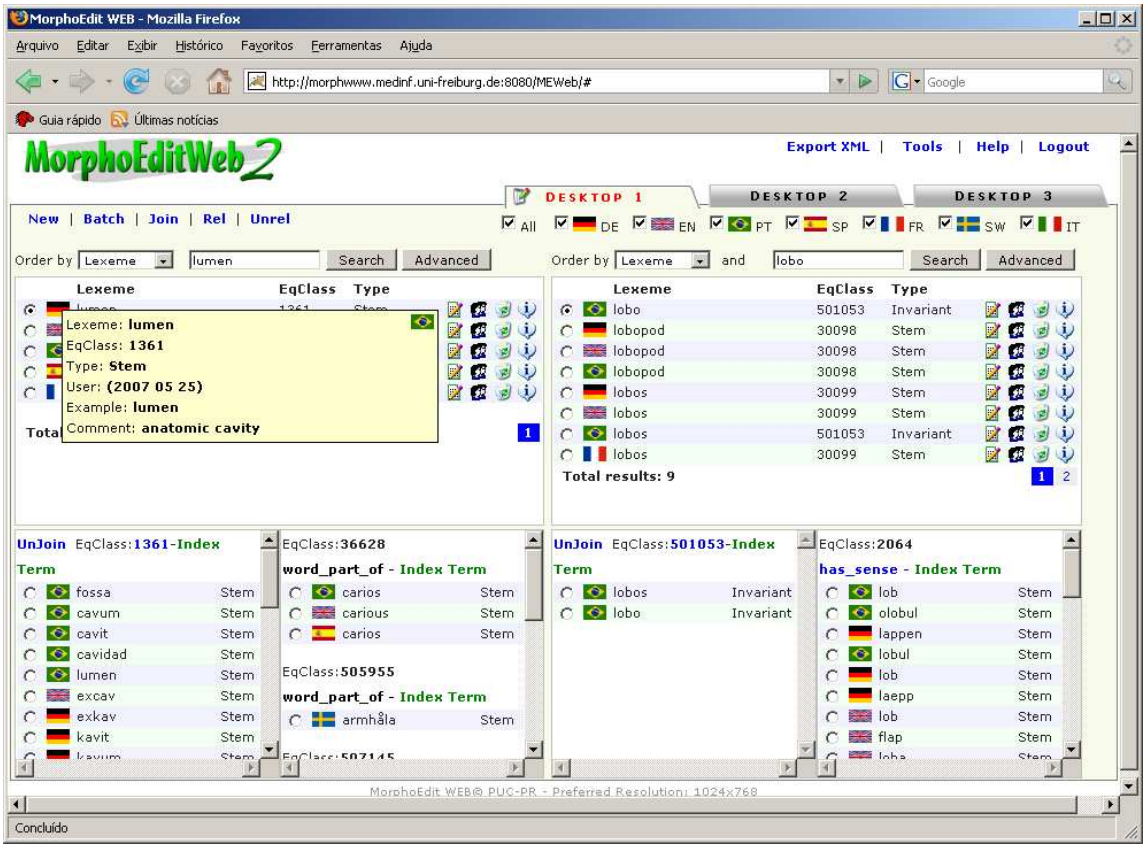



Figura 11: *MorphoEditWeb*: gerenciador do tesauro via *Web*.

Fonte compilada do UMLS

O advento de fontes de terminologia médica, tais como o *metathesaurus* do Sistema Unificado de Linguagem Médica (*Unified Medical Language System* - UMLS) da U.S. National Library of Medicine (NLM)[10] (ZHANG, RODERER, HUANG e ZHAO, 2006), ajudou no desenvolvimento de programas de indexação automática de textos biomédicos, mapeando palavras, termos e frases em textos biomédicos para conceitos. Esse *metathesaurus* é composto por centenas de milhares de conceitos médicos e suas variações léxicas são provenientes de dezenas de vocabulários controlados. Além disso, sofrem constantes atualizações referentes aos setores da área médica.

[10] http://www.nlm.nih.gov/

O UMLS foi criado para ajudar os profissionais da saúde e pesquisadores a recuperar e integrar informações biomédicas contidas em diferentes fontes. Pode ser usado para superar variações na maneira que conceitos similares são expressos em fontes diferentes. Isto torna mais fácil para os usuários ligarem a informação dos sistemas do registro dos termos.

O UMLS constitui-se de três fontes integradas que visam a normalização dos termos registrados, são elas: *Methathesaurus*, Léxico Especialista e a Rede Semântica (UMLS, 1994).

O *Methathesaurus* está organizado no conceito e no significado dos termos fazendo um *link* de nomes alternativos e visões do mesmo conceito. Ou seja, contém as equivalências, significado e as relações semânticas dos termos registrados. No especialista léxico há informações sintáticas, morfológicas e ortográficas das palavras. Apresenta as variáveis sintáticas dos termos. E na rede semântica encontram-se informações sobre os tipos e as categorias das palavras.

Quando um termo é registrado, este passa pelas três fontes, para que desta forma sejam esgotadas todas as possibilidades de uso deste termo. Para a palavra ombro, por exemplo, será encontrada no *Methathesaurus*, no qual estarão o seu significado, suas equivalências em outros idiomas, e as relações semânticas que possam existir em relação a ela, tais como: síndrome do ombro, doença em ombro, síndrome do impacto.

O *MorphoSaurus* integra um subconjunto do UMLS *Metathesaurus*, selecionado por dois critérios. Só foram incluídos termos não compostos e foram excluídos todos os idiomas que não tem relevância para o *Morphosaurus*. A maior utilidade dessa ferramenta no trabalho lexicográfico é verificar relações de sinonímia. Vale ressaltar que um grande número de palavras no *Metathesaurus* ocorre exclusivamente em termos complexos e não é incluído na lista.

Fonte compilada do MesH

O *Medical Subject Headings* (MeSH) foi criado pela NLM para ser o vocabulário de referência usado na indexação de artigos, catalogação de livros, e na busca de coleções médicas digitais, tais como a MEDLINE. O vocabulário MeSH provê uma forma consistente de recuperar informação já que é bastante detalhada com diferentes descrições para um mesmo conceito. Além disso, o MeSH organiza seus descritores em uma estrutura hierárquica tal que categorias mais abrangentes podem recuperar artigos indexados com categorias mais restritas (HEARST, 1999). Nos níveis mais abrangentes da hierarquia, encontram-se conceitos

tais como Anatomia e Distúrbios Mentais. Nos mais específicos, conceitos como Tornozelo e Distúrbio de Conduta.

O vocabulário MeSH é continuamente atualizado por especialistas de diversas áreas. A cada ano, centenas de novos conceitos são adicionados e milhares de modificações realizadas.

O MeSH é estruturado em 15 categorias hierárquicas, ou ramos da árvore do conhecimento em ciências da saúde: Anatomia (A), Organismos (B), Doenças (C), Compostos Químicos e Drogas (D), Técnicas e Equipamentos (E), Psicologia e Psiquiatria (F), Ciências Biológicas (G), Ciências Físicas (H), Antropologia, Educação, Sociologia e Fenômenos Sociais (I), Tecnologia e Alimentos e Bebidas (J), Humanidades (K), Ciências da Informação (L), Pessoas (M), Assistência à Saúde (N) e Localizações Geográficas (Z) (TARDELLI, ANCAO, PACKER e SIGULEM, 2002).

No *MorphoEditWeb*, a lista compilada tanto do UMLS quanto do MeSH Utiliza-se somente termos sinônimos (intra e multilingual) para a verificação de termos ambíguos, conforme mostra a figura 12.

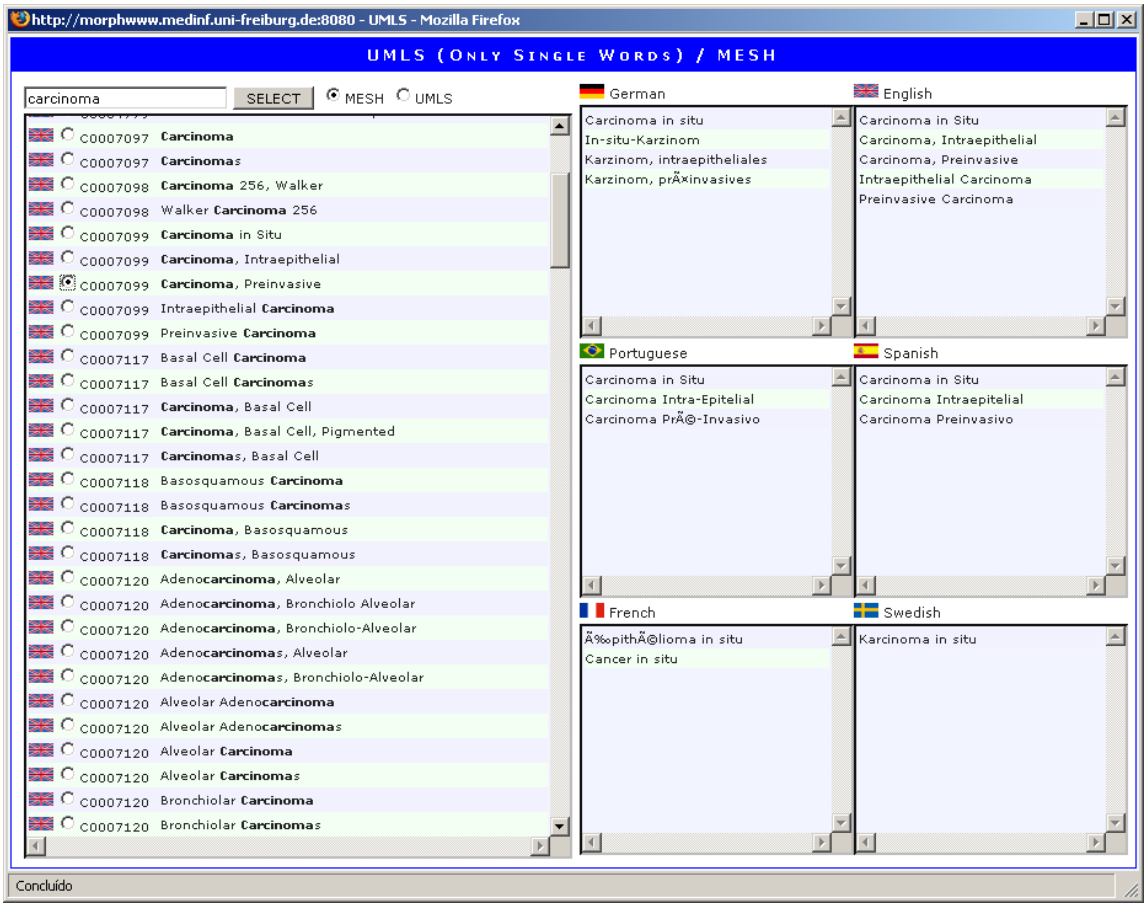



Figura 12: *MorphoEditWeb* com interfaces com fontes compilada do MeSH e UMLS.

- *Wordstat*

O recurso *Word Stat* conforme figura 15 apresenta dados estatísticos de distribuição de palavras que foi compilada a partir de *corpora* de referência extraídas da *Web* (especialmente, do Manual MSD, adotado em vários idiomas [...]). O usuário pode pesquisar a lista por *substring* e ordenar ou por ordem alfabética ou por ordem de freqüência. *Wordstat* auxilia o lexicógrafo ao recuperar todas as palavras que incluem uma *substring* candidata a lexema. Isso importa especialmente com *strings* curtas de três ou quatro caracteres que, às vezes, ocorrem em múltiplos contextos.

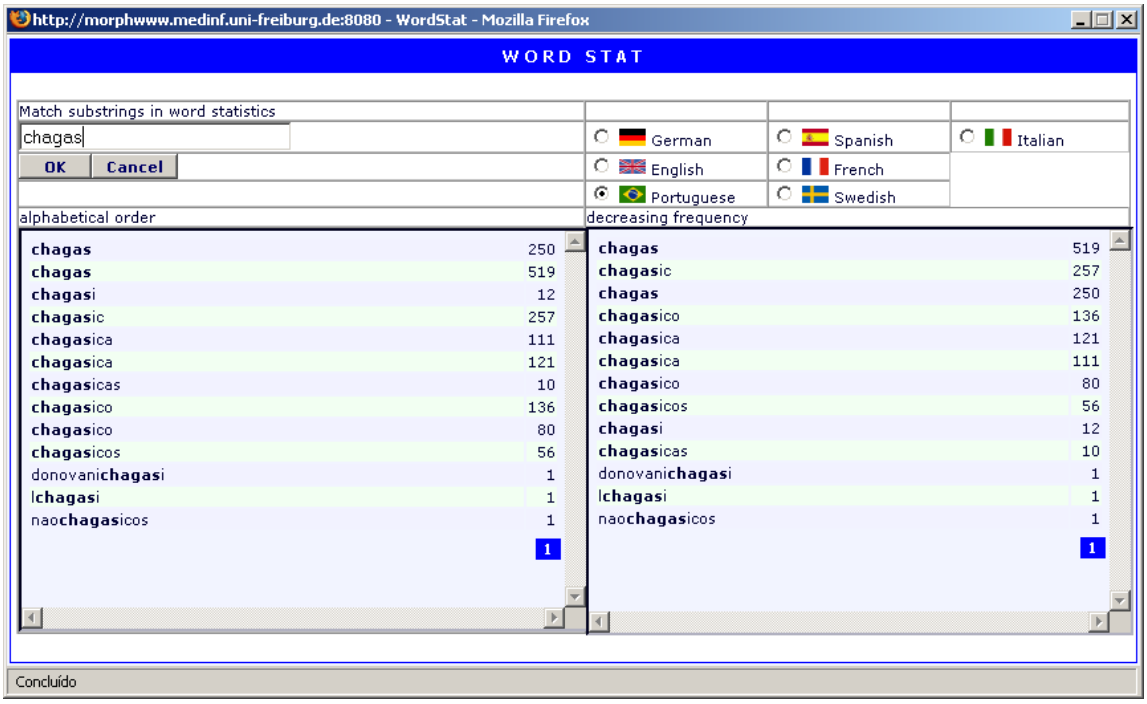



Figura 13: *MorphoEditWeb* e a ferramenta de apoio Wordstat.

2.10.5.2 Dados Estatísticos do Tesauro

A construção manual do tesauro do Projeto *Morphosaurus*, consumiu até então quatro anos de trabalhos. Os trabalhos começaram inicialmente com inglês e alemão, vindo, então, a ser incluído o português. O léxico português carece da terminologia de fármacos e drogas.

Nas outras línguas, as entradas referentes às drogas foram compiladas do UMLS e do MesH. Outras entradas foram incluídas de forma automática – utilizando-se de técnicas de *bootstrapping* (SCHULZ, MARKÓ, HAHN *et al.*, 2004), no espanhol, sueco e francês, mas que ainda precisam ser validadas.

Atualmente, o tesauro conta com 90.550 entradas lexicais, com 23.976 para o léxico alemão, 22.561 para o léxico inglês, 14.984 para o léxico em português, 10.936 para o léxico espanhol, 7.812 para o léxico francês e 10.281 para o léxico sueco. Todas as entradas estão relacionadas por 21.432 classes de equivalências. O mesmo possui uma cobertura consolidada para os léxicos em inglês, português e alemão. Os léxicos em sueco, espanhol e francês continuam em fase de construção.

### 2.10.6 A Segmentação pelo Sistema *Morphosaurus*

O léxico e tesauro manipulado pelos lexicógrafos são armazenados numa base de dados MySQL. Porém, para verificar o resultado de uma normalização morfossemântica de uma palavra ou uma lista de palavras, ou de uma página HTML, o operador do léxico deve exportar o léxico e tesauro para o padrão XML para que as ferramentas *Morphosaurus* possam processá-las.

Devido às peculiaridades e regras gramaticais de cada língua, cada idioma possui o seu próprio segmentador, cujas regras são implementadas em arquivos separados. Essa ferramenta é muito utilizada para verificar visualmente o resultado da segmentação de uma palavra, uma lista de palavras ou uma amostra de texto, inclusive de forma bilíngüe. A figura 14 mostra a tela de entrada de dados a ser segmentada e a figura 15 mostra o resultado devolvido pelo segmentador da língua escolhida no sistema *Morphosaurus*.

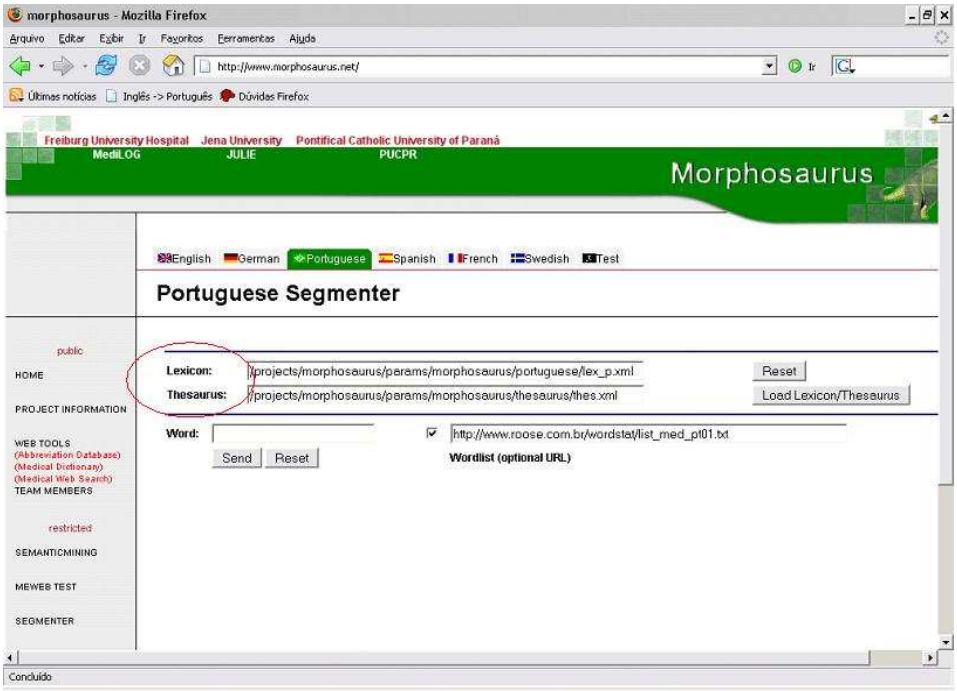
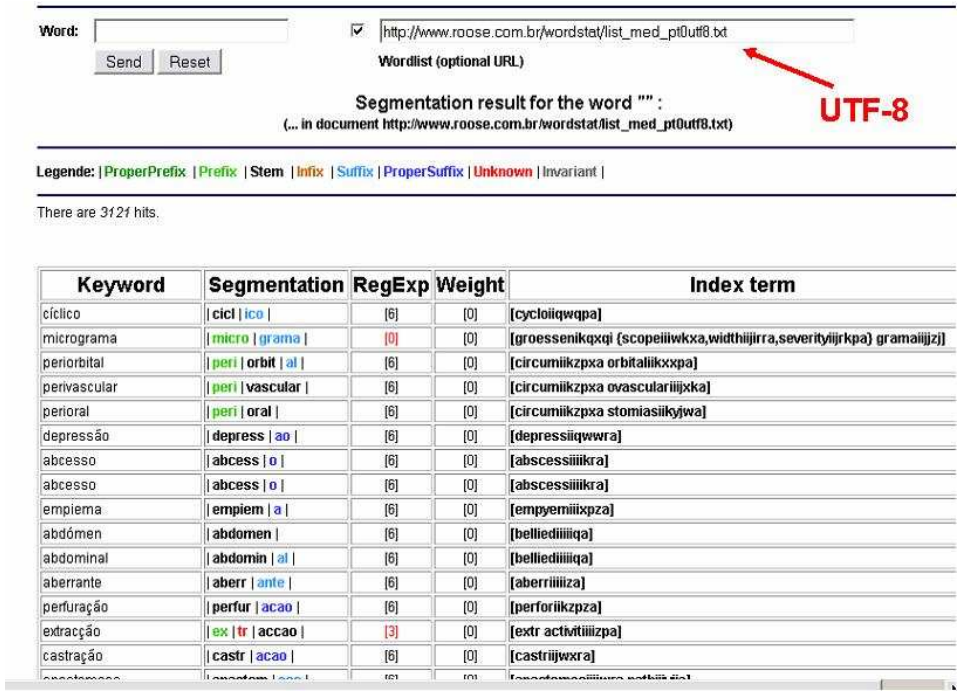



Figura 14: Interface do Módulo Segmentador do Sistema *Morphosaurus*.




| Keyword | Segmentation | RegExp | Weight | Index term |
|---|---|---|---|---|
| cíclico | \| cicl \| ico \| | [6] | [0] | [cycloiiqwqpa] |
| micrograma | \| micro \| grama \| | [0] | [0] | [groessenikqxqi {scopeiiiwkxa,widthiijirra,severityiijrkpa} gramaiijjzj] |
| periorbital | \| peri \| orbit \| al \| | [6] | [0] | [circumiikzpxa orbitaliikxxpa] |
| perivascular | \| peri \| vascular \| | [6] | [0] | [circumiikzpxa ovasculariiijxka] |
| perioral | \| peri \| oral \| | [6] | [0] | [circumiikzpxa stomiasiikyjwa] |
| depressão | \| depress \| ao \| | [6] | [0] | [depressiiqwwra] |
| abcesso | \| abcess \| o \| | [6] | [0] | [abscessiiiikra] |
| abcesso | \| abcess \| o \| | [6] | [0] | [abscessiiiikra] |
| empiema | \| empiem \| a \| | [6] | [0] | [empyemiiixpza] |
| abdómen | \| abdomen \| | [6] | [0] | [belliediiiiqa] |
| abdominal | \| abdomin \| al \| | [6] | [0] | [belliediiiiqa] |
| aberrante | \| aberr \| ante \| | [6] | [0] | [aberriiiiza] |
| perfuração | \| perfur \| acao \| | [6] | [0] | [perforiikzpza] |
| extracção | \| ex \| tr \| accao \| | [3] | [0] | [extr activitiiizpa] |
| castração | \| castr \| acao \| | [6] | [0] | [castriijwxra] |

Figura 15: lista de palavras segmentadas e normalizadas (MIDs).

# CAPÍTULO 3

# METODOLOGIA

## 3.1 MATERIAIS E INFRA-ESTRUTURA

A idéia principal desse trabalho é criar uma metodologia que aponte potenciais erros no conteúdo lexical e semântico do Sistema *Morphosaurus*. Esses problemas podem ser oriundos de classes de equivalências mal delimitadas, relacionamentos errôneos do tesauro, assim como também da própria heurística de segmentação implementada nos módulos de segmentação e indexação do sistema baseado na abordagem por *subwords*. Para realização de tais tarefas, foi necessário montar um *workbench*.

De forma macro, as seguintes ferramentas foram necessárias:

(1) um sistema de indexação morfossemântica multilíngüe baseada em tesauro multilíngüe;
(2) um motor de busca - de onde se utilizou os módulos de indexação, para geração dos arquivos invertidos;
(3) uma coleção de teste como padrão ouro para recuperação de documentos do domínio médico;
(4) corpora comparáveis multilíngüe do domínio médico;
(5) rotinas de computador para processamento lingüístico e geração de curvas de precisão e revocação.

Um trabalho realizado com uma equipe multidisciplinar composta por usuários de lugares e línguas diferentes exige um ambiente computacional para suportar tal cenário heterogêneo. Pelo fato de se trabalhar com os parceiros alemães da Universidade de Freiburg, optou-se por criar uma infra-estrutura computacional de serviços compatível com a rede de servidores e computadores desta universidade, de forma a suportar alguns serviços e conectividade com esta de forma estável. Para isso, foram implementadas inicialmente (alguns softwares foram atualizados posteriormente) as seguintes ferramentas e configurações para execução de processos e verificação de resultados:

**Softwares**

Ambiente *Unix / Linux*

*Java 2 Plataform, Standard Edition Version 1.4.0*

*Perl 5.6.1*

*Apache HTTP Webserver 1.3.26*

*HtDig 3.1.6*

*MySQL*

**Serviços**

*WWW – World Wide Web*

Protocolos *TCP/IP, HTTP, HTTPS*

SSH

Diretório (Acesso aos arquivos)

*URL*: http://www.ler.pucpr.br/~roose

**Equipamentos**

Um Servidor *Xeon Dual Processor* foi adquirido para os propósitos descritos. A configuração do equipamento foi de tal maneira a ter escalabilidade no armazenamento de informações, haja vista a incerteza na quantidade de documentos que viria a ser utilizado no trabalho.

*Configuração Física e Diagrama de Conexões dos Servidores*

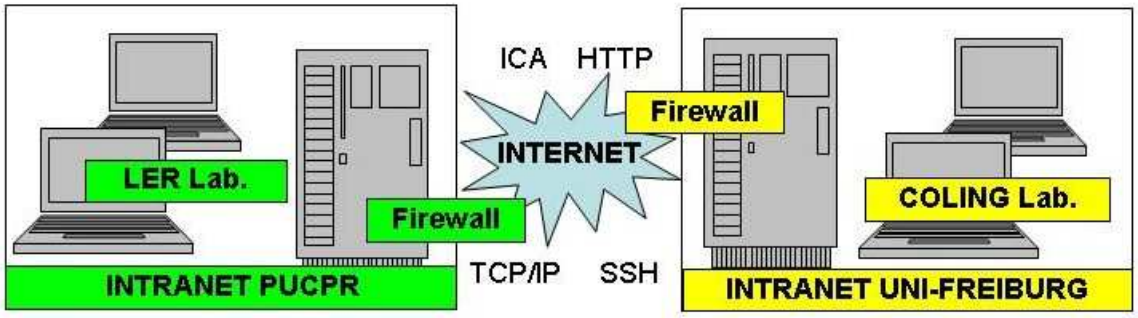



Figura 16: Diagrama de conexão de rede entre a PUCPR e a Uni-Freiburg.

**Linguagem de Programação**

Foram utilizadas diversas ferramentas computacionais. Considerando as plataformas empregadas, houve necessidade de utilizar linguagens de programação tanto em Windows quanto em Linux. Para isso, levou-se em conta a adequação da ferramenta à tarefa solicitada e sua facilidade de implementação. As rotinas de normalização foram programadas na linguagem JAVA e, para a execução destas, foram implementadas rotinas na linguagem PERL e *Shell Script*.

*Fatores que motivaram o uso da linguagem JAVA*

A linguagem Java, criada pelo grupo liderado por James Gosling na *Sun Microsystems*, é uma linguagem computacional completa, independente de plataforma e com uma série de facilidades para a integração com a Internet (SUN, 1995).

Os fatores que motivaram o uso da linguagem Java do ponto de vista Computacional foram:

(1) multi-plataforma: o compilador Java compila o código Java em "*bytecodes*". Estes *bytecodes* são, então, interpretados por uma "Máquina Virtual" Java, que é escrita para a arquitetura de processador em que o programa virá a rodar, isto permite funcionar em qualquer sistema operacional;
(2) linguagem Orientada a Objetos, permitindo a reutilização de código, assim aumentando a produtividade.
(3) *Java Database Connectivity* – JDBC: utilizada para acesso ao banco de dados. Trabalha em conjunto com o *driver* do banco de dados. É utilizada para todas as funções como consultas, inclusão e exclusão de registros (SUN, 2001);
(4) utilização do *Java Server Pages* – JSP: tecnologia baseada em Java que simplifica o processo de desenvolvimento de sites dinâmicos. JSP é composto de tags, que são incluídas junto ao código HTML para serem executadas durante uma requisição. O código JSP é compilado para Java e isto garante melhor desempenho do que linguagem de scripts interpretados (SUN, 2001a).

Os fatores que motivaram o uso da linguagem Java do ponto de vista lingüístico foram:

(1) Java é muito adequado para trabalhar com linguagem natural devido a sua avançada capacidade em lidar com conjunto de caracteres;
(2) Java suporta totalmente a codificação *Unicode*[11]. Essa codificação é capaz de operar com alfabetos não-latinos como grego, cirílico e, até mesmo, o chinês.

Para o tráfego de informações das fontes de dados utilizou-se o padrão XML . A figura 17 apresenta um terminal com os diretórios de ferramentas, *scripts* e sub-diretórios padronizados com uma versão de tesauro.

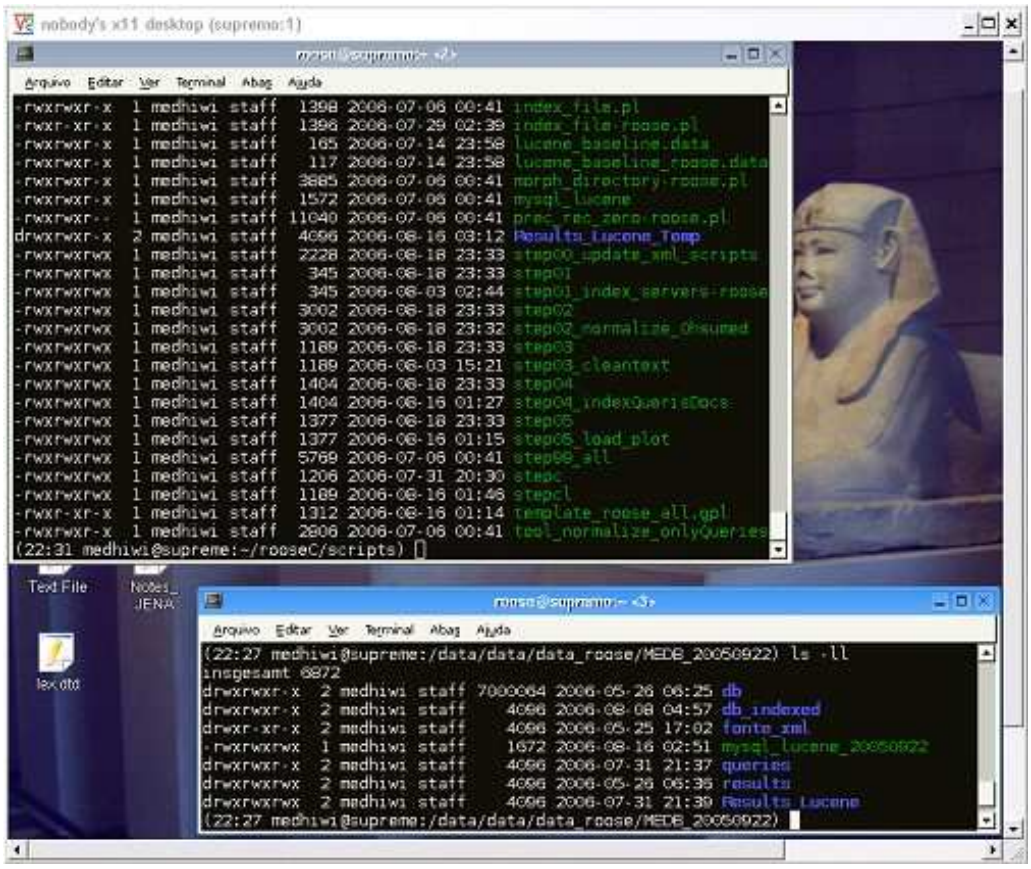



Figura 17: Acesso remoto aos servidores da Uni-Freiburg para execução de *scripts*.

[11] http://www.unicode.org

## 3.2 DESENVOLVIMENTO

Basicamente, os procedimentos descritos foram realizados seqüencialmente:

(1) montagem de *corpora* do domínio médico para fins estatísticos (inglês, alemão, português, espanhol e sueco);
(2) normalização morfossemântica desses *corpora*;
(3) geração de listas de freqüências bilíngües nos idiomas propostos;
(4) verificação e correção de classes suspeitas conforme lista de freqüência;
(5) acompanhamento do processo de correção através repetidas medições de parâmetros de desempenho (precisão e revocação) em experimentos de recuperação de informação, usando um padrão ouro existente.

Poder-se-ia resumir os procedimentos descritos pela sua complexidade ou demora, em três grupos: (a) geração das listas de freqüências de MIDs com os procedimentos 1-2-3, (b) trabalho de lexicografia com o procedimento 4 e (c) verificação do desempenho com os procedimentos 5.

### 3.2.1 Montagem de Corpora MSD

As fontes textuais para a montagem de *corpora* nas línguas inglesa, alemã, portuguesa, espanhola e sueca, foram obtidas do *site* da Merck[12]. Depois de longa e árdua procura, decidiu-se escolher essas fontes (chamadas MSD manual[13]) porque seu conteúdo representava o mesmo assunto em outras línguas dando a característica de fontes de textos comparáveis. Não existindo esse recurso na língua sueca, optou-se pelo *corpus* do *site* Netdoktor[14]. Não obstante tal exceção, essa coleção de *corpora* será referenciada pelo nome *coleção MSD*.

Com o objetivo de criar a distribuição de ocorrências entre os identificadores semânticos no sistema *Morphosaurus* (MIDs) entre os *corpora* multilíngüe, utilizou-se a coleção MSD que foi submetido ao indexador *Morphosaurus* para a geração das tabelas de ocorrências para cada língua.

---

[12] http://www.merck.com
[13] *Merck Sharp & Dohme Manual of Clinical Medicine, disponível em inglês, português, alemão e espanhol*
[14] http://www.netdoktor.se

Textos bilíngües existem de várias formas. Podem ser paralelos ou comparáveis. Textos paralelos são aqueles para os quais os textos bilíngües possuem tradução mútua. O problema é que na tradução, um texto traduzido pode não expressar a informação do texto fonte tornando sua montagem difícil mesmo que restrito a um domínio. Esses são chamados de textos paralelos ruidosos. *Corpora* Comparáveis são aqueles que possuem amostras de textos em pares bilíngües que podem ser comparados por possuir características pré-definidas comuns entre eles como, por exemplo, o domínio, tópico, autores, etc... (DÉJEAN, GAUSSIER e SADAT, 2002).

A abordagem aqui proposta está baseada na suposição que há uma correlação entre a freqüência de ocorrência das palavras no *corpus* em um idioma A comparado com a freqüência de ocorrência das traduções correspondentes no *corpus* comparável de um idioma B. (RAPP, 1995; FUNG, 2000). É de se esperar que a distribuição de descritores semânticos (como as MIDS do *Morphosaurus*) em cada *corpus* exiba um alto grau de conformidade. Se houver uma discrepância muito grande de distribuição de descritores semânticos (MIDs), então isso pode ser um indício de o termo estar com algum tipo de problema no tesauro.

O acesso às fontes foi autorizado mediante um termo de responsabilidade para o uso das informações que, nesse caso, restringiu-se somente a gerar um grande arquivo com comportamento de *corpus*. Para a montagem do *corpus* sem tratamento do conteúdo, utilizou-se de ferramentas nativas do Linux.

Para cada idioma, os textos tiveram parágrafos duplicados removidos e foram armazenados num único arquivo sem nenhum processamento estatístico.

### 3.2.2 Normalização de cada *Corpus* Estatístico MSD

Para a geração das listas de freqüências das MIDs em cada *corpus* MSD, foi necessário processar os *corpora* com as ferramentas do *Morphosaurus* para realizar a normalização morfossemântica de cada *corpus* dos idiomas. Para isso, foram realizados os seguintes procedimentos:

(1) normalização morfossemântica dos corpora MSD com base na tesauro do mês de julho de 2005;
(2) geração de cópias de segurança (*backups*) diários do tesauro;

(3) desenvolvimento de rotinas para converter o tesauro para o padrão XML conforme a especificação do sistema *Morphosaurus*, com o objetivo de comparar diferentes versões do tesauro com relação ao seu desempenho em cenários de recuperação de informação.

### 3.2.3 Geração das listas de ocorrências das MIDs bilíngües

Depois do processamento morfossemânticodo *corpora* MSD pelo indexador do Sistema *Morphosaurus* gerou-se listas com as freqüências de cada MID em cada idioma. O objetivo foi confrontar, de forma bilíngüe, as MIDs das listas e priorizar aquelas com maior discrepância de ocorrências. Para facilitar a classificação das MIDs concorrentes, decidiu-se gerar um índice (*score* - S) que expressa a ocorrência de uma MID numa determinada língua com relação à outra. Esse índice foi parametrizado, de acordo com as equações (6), (7) e (8) para ficar entre 0 (zero) e 1 (um); onde as MIDs, próximo da unidade, indicam uma maior probabilidade de estar com algum tipo de problema em potencial[15]. Para cada descritor da lista, verificou-se sua real correspondência ao seu significado; caso fosse detectada alguma inconsistência, uma correção nessa era realizada através do gerenciador *MorphoEditWeb*. As tabelas 2 e 3 apresentam os primeiros descritores, classificados pelo índice S, referentes à comparação entre os idiomas português-inglês e alemão-inglês.

$$S = \frac{2S_d + S_a}{3} \tag{6}$$

$$S_d = \frac{|f1 - f2|}{|f1 + f2|} \tag{7}$$

$$S_a = \frac{fx}{(fx1 + fx2)_{\max}} \tag{8}$$

onde:

(1) $f1$ *é a freqüência da ocorrência de uma MID num corpus;*

(2) $f2$ *é a freqüência da ocorrência de uma MID em outro corpus;*

[15] A lista de comparação encontra-se disponível no link "*MIDCompare*" em http://www.ler.pucpr.br/~roose /dissertation.

(3) *fx refere-se aos índices de cada linha da lista de MIDs comparáveis (de uma língua em relação à outras);*

*(4)* $(fx1 + fx2)_{max}$ *corresponde ao valor máximo da ocorrência do descritor em cada idioma;*

(5) $S_d$ *expressa um índice com base na diferença de ocorrência de uma MID em um corpus normalizado em relação a outro;*

(6) $S_a$ *relaciona o valor relativo da ocorrência de uma MID com relação ao maior índice de ocorrência em ambas as listas;*

(7) *S é o índice final com o objetivo de mostrar indícios de problemas no tesauro normalizado entre 0 e 1.*

Para a realização dos cálculos necessários e seus resultados, conforme equações listadas, foram desenvolvidas rotinas na linguagem de programação JAVA.

### 3.2.4 Verificação e correção de classes de equivalências suspeitas

Uma vez dividida a lista com quantidades equivalentes para os lexicógrafos, começaram as devidas correções. Para que houvesse uma sincronização adequada entre os integrantes do projeto, nomeou-se um responsável pelas correções dos léxicos e tesauro das línguas alemã, sueca e inglesa; enquanto o responsável no Brasil ficou a cargo das correções entre as MID´s portuguesa, espanhola e inglês. Em alguns casos, também entre português e alemão. Um esquema geral do cenário pode ser verificado na figura 18.

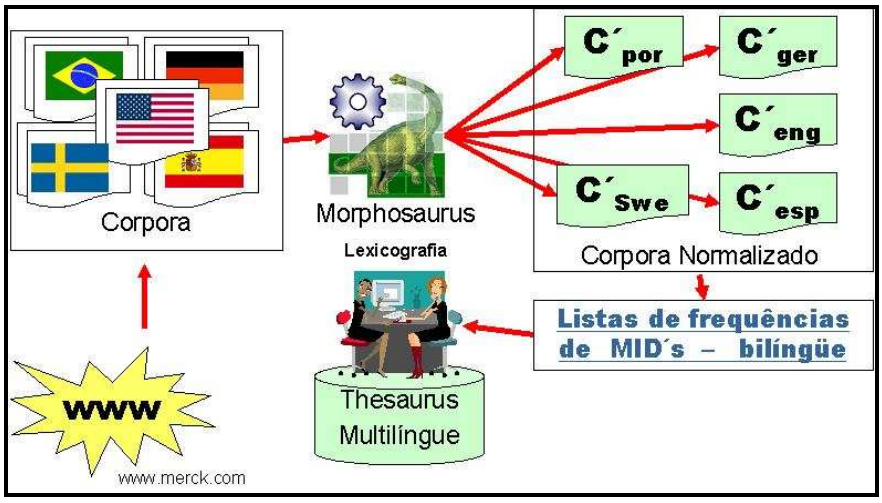



Figura 18: *Workbench* para a verificação de erros no tesauro.

Para registro das alterações realizadas no tesauro, utilizou-se de um formulário padrão (Tabela 4) que, então, era submetido a um fórum de discussão, no sentido de facilitar a comunicação sobre os erros e também como forma de suporte a difíceis decisões sobre a resolução destas. Entretanto, observou-se que muitas das decisões eram baseadas em consenso, ou seja, muitas das soluções eram resolvidas de forma não sistemática. Apesar disso, esse procedimento respondeu com muito boas respostas aos problemas expostos.

Tabela 3: Amostra de freqüências das MIDs e seus parâmetros relacionados entre português ($f1$) e inglês ($f2$) para um fx1 = 9363 e fx2 = 9369.

| **MID** | **MIDCod** | $f1$ | $f2$ | $S_a$ | $S_d$ | $S$ |
|---|---|---|---|---|---|---|
| Peopleriixypa | 500783 | 6352 | 0 | 0,1466 | 1,0000 | 0,7155 |
| Fromiwiixxa | 060077 | 4676 | 0 | 0,1079 | 1,0000 | 0,7026 |
| Icasikprrr | 023555 | 0 | 3022 | 0,0697 | 1,0000 | 0,6899 |
| Lttroriiyira | 500805 | 10 | 3331 | 0,0771 | 0,9940 | 0,6884 |
| Mostiizrpwa | 009536 | 2783 | 0 | 0,0642 | 1,0000 | 0,6881 |
| Enteikywjw | 028616 | 0 | 2069 | 0,0477 | 1,0000 | 0,6826 |
| Icakiirwy | 200568 | 0 | 1945 | 0,0449 | 1,0000 | 0,6816 |
| Sometimerijixja | 501071 | 1708 | 0 | 0,0394 | 1,0000 | 0,6798 |
| Pressureiiipkza | 000329 | 1833 | 2 | 0,0423 | 0,9978 | 0,6793 |

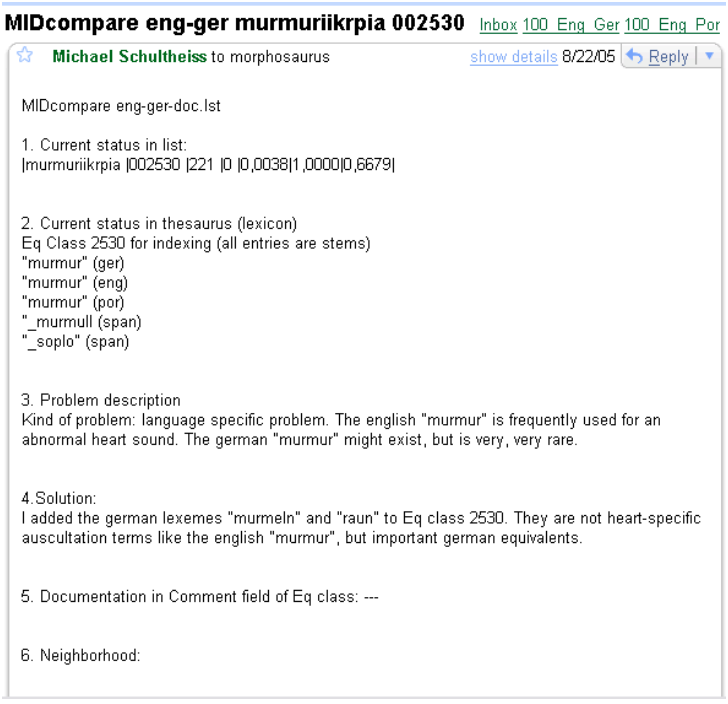
Tabela 4: Amostra de freqüências das MIDs e seus parâmetros relacionados entre alemão ($f1$) e inglês ($f2$).

| **MID** | **MIDCod** | $f1$ | $f2$ | $S_a$ | $S_d$ | $S$ |
|---|---|---|---|---|---|---|
| Zpippxra | 303375 | 1 | 3428 | 0,0590 | 0,9994 | 0,6859 |
| Keinemrikzrp | 502953 | 0 | 1803 | 0,0310 | 1,0000 | 0,6770 |
| Barriqrqp | 504543 | 0 | 1021 | 0,0176 | 1,0000 | 0,6725 |
| eingesetztijiikr | 010025 | 0 | 972 | 0,0167 | 1,0000 | 0,6722 |
| Ipippry | 303358 | 0 | 956 | 0,0165 | 1,0000 | 0,6722 |
| dispensatrijiyya | 501088 | 0 | 845 | 0,0145 | 1,0000 | 0,6715 |
| langerrikzzwa | 502996 | 0 | 780 | 0,0134 | 1,0000 | 0,6711 |
| Siterijjrka | 501152 | 681 | 0 | 0,0117 | 1,0000 | 0,6706 |

Tabela 5: Formulário para registro de alterações no tesauro pelos os lexicógrafos.

| ***MIDcompare*** |
|---|
| *1. Current status in list* |
| *2. Current status in thesaurus (lexicon)* |
| *3. Problem description and kind of problem* |
| *4. Solution and Reasons* |
| *5. Documentation in Comment field of Eq class* |
| *6. Neighborhood* |
| *7. Open questions / to do* |

Inicialmente, cada caso verificado foi comentado, resolvido e justificado. Os casos sem solução também passaram pelo mesmo processo até um consenso final. Depois de uma quantidade razoável de correções, adotou-se consultar a MID ou o seu número da classe de equivalência no fórum para verificar se a mesma já havia sido alterada. As figuras 19 e 20 apresentam um exemplo típico de um formulário preenchido, inglês-alemão e português-espanhol respectivamente.



**MIDcompare eng-ger murmuriikrpia 002530** Inbox 100 Eng Ger 100 Eng Por

**Michael Schultheiss** to morphosaurus show details 8/22/05 Reply

MIDcompare eng-ger-doc.lst

1. Current status in list:
|murmuriikrpia |002530 |221 |0 |0,0038|1,0000|0,6679|

2. Current status in thesaurus (lexicon)
Eq Class 2530 for indexing (all entries are stems)
"murmur" (ger)
"murmur" (eng)
"murmur" (por)
"_murmull (span)
"_soplo" (span)

3. Problem description
Kind of problem: language specific problem. The english "murmur" is frequently used for an abnormal heart sound. The german "murmur" might exist, but is very, very rare.

4.Solution:
I added the german lexemes "murmeln" and "raun" to Eq class 2530. They are not heart-specific auscultation terms like the english "murmur", but important german equivalents.

5. Documentation in Comment field of Eq class: ---

6. Neighborhood:

Figura 19: Protocolo de comunicação entre lexicógrafos – inglês e alemão.

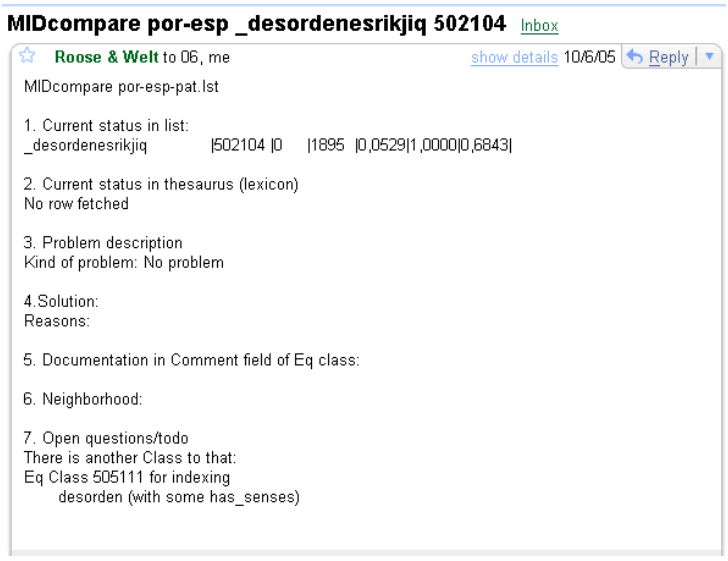



Figura 20: Protocolo de comunicação entre lexicógrafos – português e espanhol.

### 3.2.5 Avaliação do tesauro

Inicialmente, decidiu-se analisar os 100 primeiros MIDs dos idiomas entre inglês-português, inglês-alemão e português-espanhol. Mas, pela falta de sincronismo entre horários (e fuso-horário) dos lexicógrafos e com o intuito de melhorar a confiabilidade dos resultados finais, optou-se por analisar as 160 MIDs, aproximadamente, corrigidas no período de 3 meses, totalizando em torno de 100 h por lexicógrafo.

A avaliação do tesauro partiu da seguinte hipótese: considerando um tesauro como um componente dentro de um sistema de recuperação de informação, avaliar a qualidade de um tesauro, indiretamente é avaliar o desempenho de um sistema de recuperação de informações. A qualidade de um tesauro está diretamente ligada em sua resposta como uma fonte de conhecimento no domínio médico de forma que seja satisfatória na geração de descritores semânticos em conformidade com seus reais sentidos, inclusive na geração de suas possíveis acepções de forma também normalizada. Desta forma, assume-se que a boa qualidade de um tesauro irá refletir de forma positiva no desempenho de um sistema de recuperação de informações que o utiliza.

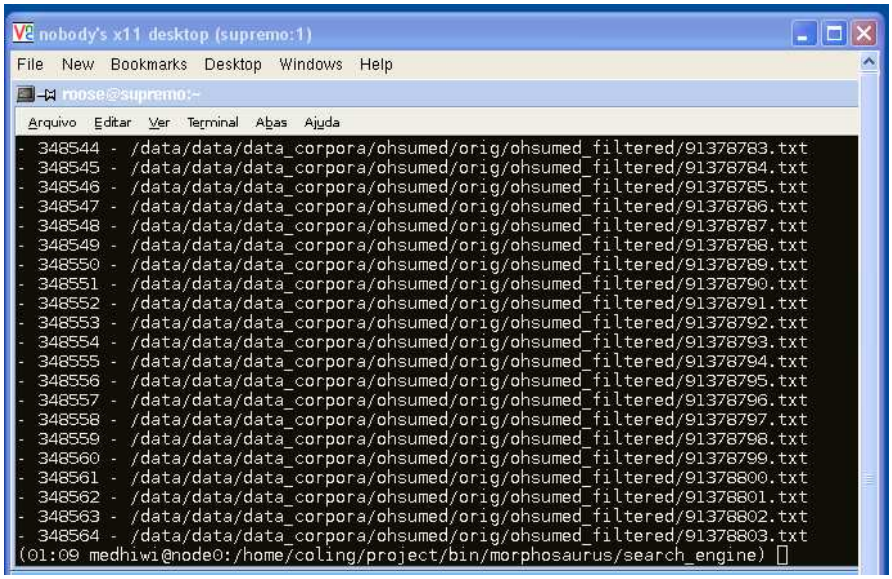



Figura 21: Final de processo de normalização morfossemântica da coleção OHSUMED.

Existem várias técnicas e formas avaliar um sistema de recuperação de informações (BAEZA-YATES e RIBEIRO-NETO, 1999). Nesse trabalho, para avaliação do desempenho da proposta, utilizou-se a coleção de teste OHSUMED baseado em uma seleção de sumários MEDLINE[16] (HERSH, 1996) como padrão ouro para levantar as curvas de precisão e revocação (*precision vs. recall*) para análises e verificações de ganho ou perda de desempenho. Resumiu-se o desempenho da metodologia proposta através da evolução das médias dos valores de precisão sobre os onze pontos de revocação com base no período de correções definido.

As curvas de precisão e revocação foram plotadas com intervalos de sete dias, a partir do primeiro *backup* tomado como oficial para esse trabalho. Foram remontados 10 tesauros com os *backups* dos dias 02/08/2005, 09/08/2005, 16/08/2005, 23/08/2005, 01/09/2005, 08/09/2005, 15/09/2005, 22/09/2005, 01/10/2005 e 08/10/2005.

A seguir, apresenta-se a seqüência de procedimentos para levantar uma curva de precisão e revocação para qualquer versão de tesauro do sistema *Morphosaurus*.

Primeiramente, levantou-se a curva de referência com base nas *queries* e documentos originais em inglês, a *baseline*, para poder comparar com as curvas de outros idiomas. Mas, por questões de praticidade, primeiramente será explicado como se plotam curvas de precisão e revocação nesse trabalho com todos os documentos normalizados; pois, para montar a

[16] http://www.nlm.nih.gov/research/umls/umlsmain.html

*baseline*, bastará excluir um dos procedimentos. Esses procedimentos estão separados em blocos conforme segue:

COLEÇÃO DE TESTE OHSUMED

1) Estudantes de medicina e médicos bilíngües traduziram os *queries* (originais em inglês) para os idiomas português, alemão, espanhol (a língua sueca também foi incluída automaticamente no processo, mas não foi objeto de análise nesse processo);
2) Filtraram-se os documentos da coleção sem conteúdo textual do campo resumo reduzindo-o para 233.445 documentos (67%).

SERVIDORES

3) Conferiu-se o tipo de codificação utilizado pelo servidor da Alemanha: sistema UNIX, e padrão de codificação UTF-8;
4) Conferiu-se as portas de conexão necessárias envolvidas pelos servidores de conexão e módulos de normalização do sistema *Morphosaurus*.

SISTEMA *MORPHOSAURUS*

Restaurou-se a versão de léxico de interesse no banco de dados MySQL e, a partir deste, gerou-se o tesauro no padrão XML. Armazenaram-nos em diretórios adequados e ajustaram-se as configurações necessárias e exigidas pelos módulos servidores de indexação e segmentação:

5) reiniciaram-se os servidores indexadores (*index_server*) de cada idioma do sistema *Morphosaurus*;
6) executou-se s*cript* para normalizar *queries* e documentos normalizados;
7) executou-se *script* para indexar documentos;
8) executou-se *script* para plotar as curvas de precisão e revocação.

Para a geração da *baseline*, basta ajustar os arquivos de configuração para normalizar as *queries* e os documentos da coleção OHSUMED sem normalizar, ou seja, os procedimentos são realizados com as *queries* e documentos no original em inglês - não se executa o passo 6.

Os dados de arquivos invertidos e outros parâmetros gerados por programas específicos foram armazenados no banco de dados. Depois foram executadas outras rotinas

em PERL para realizar os cálculos de precisão e revocação e os dados também foram armazenados no banco de dados.

Pelo fato de se tratar de vários procedimentos trabalhosos e demorados (empregou-se em torno de 5 h para normalizar os documentos da coleção OHSUMED para cada versão de tesauro), houve necessidade de escrever alguns *scripts* para automatizar algumas tarefas para maior segurança e o bom andamento seqüencial das tarefas.

Na figura 24, apresenta-se o esquema seqüencial dos procedimentos para geração das curvas de precisão e revocação para avaliação da abordagem deste trabalho. As *queries* estão representadas pelos blocos $Q^{por}$, $Q^{ger}$, $Q^{eng}$, $Q^{spa}$, $Q^{swe}$. Os procedimentos para montar o arquivo invertido de uma coleção, normalizada ou não, está representada pelo bloco "Máquina de Busca". Após esse procedimento, foram realizados os cálculos e gerados dados para montagem das curvas; que para tal utilizou-se a ferramenta *Gnuplot*.

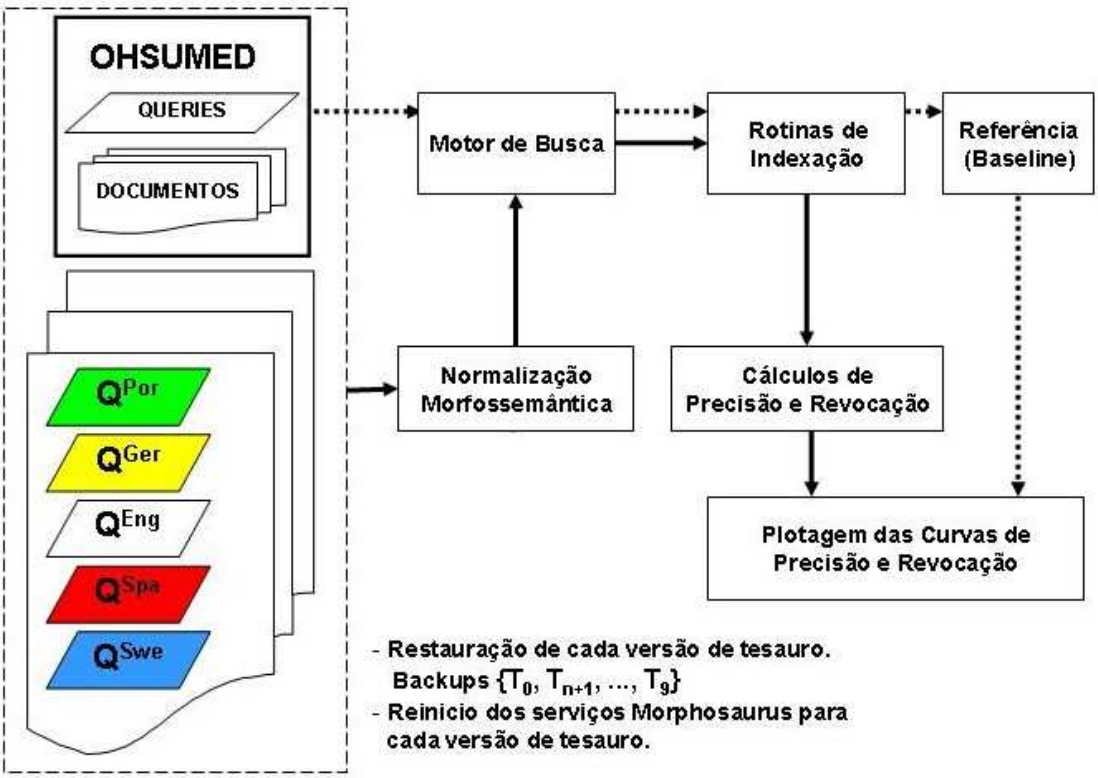



Figura 22: Esquema para avaliação do tesauro com a técnica de Precisão e Revocação.

# CAPÍTULO 4

# RESULTADOS

Num primeiro momento, foram gerados os subsídios necessários aos lexicógrafos. Estes, por sua vez, seguiram os procedimentos descritos na metodologia onde o foco se restringiu em seguir uma lista organizada por índices numéricos representando potenciais MIDs com problema, registrando as alterações e manipulações através de um protocolo definido pela equipe. Paralelamente, efetuou-se *backups* das versões do léxico com o objetivo de verificar a evolução da qualidade do tesauro através da análise cronológica das curvas de precisão e revocação tomando como referência o padrão ouro da coleção de teste OHSUMED.

No início, restringiu-se em seguir a lista de freqüência bilíngüe e registrar as alterações e manipulações. Não era possível ainda categorizar os tipos de problemas. Após o término das correções (três meses), foi possível categorizar os tipos de problemas conforme descrito na tabela 6. Restringiu-se a categorização dos problemas através daqueles encontrados nas MIDs com base na tríade inglês-português, inglês-alemão e inglês-espanhol.

## 4.1 TIPOS DE PROBLEMAS ENCONTRADOS

Categorizar os tipos de problemas foi outro fator complicador neste trabalho, pois alguns casos possuíam mais de uma categoria. Um dos fatores de dificuldade na correção das MIDs ocorreu devido à falta de exemplos nas línguas nativas do lexema criado. A tabela 6 resume os principais problemas encontrados durante as correções.

Os problemas de ambigüidade decorrem, principalmente, dos lexemas ambíguos (na mesma classe de equivalência) numa língua que não era o caso em outro idioma. Em alguns casos, foi necessário separar tal grupo de lexemas numa outra classe de equivalência e mapeá-los para as suas devidas acepções para não haver conflito entre os sentidos e idiomas. Esse procedimento foi realizado pela relação do tipo "*has_sense*". Nesse grupo também estão relacionadas classes "órfãs", ou seja, classes ambíguas sem seus respectivos relacionamentos semânticos.

Tabela 6: Problemas identificados durante as correções das MIDs.

| Tipos de Problemas | Freqüência Português / Inglês | Freqüência Alemão / Inglês | Freqüência Espanhol / Inglês |
|---|---|---|---|
| Ambigüidades | 0.23 | 0.38 | 0.14 |
| Sem entradas / dispensas | 0.49 | 0.18 | 0.53 |
| MIDs diferentes com mesmo sentido | 0.06 | 0.12 | 0.19 |
| Mesma MID com sentidos diferentes | 0.04 | 0.05 | 0.06 |
| Sem problema | 0.11 | 0.10 | 0.04 |
| Sem classificação | 0.07 | 0.17 | 0.04 |

Alguns casos de falta de classe de equivalência e falta de termos dentro da classe foram encontradas. Os casos de dispensa referem-se aos casos em que, para um mesmo sentido, eram considerados relevantes para indexação numa língua, mas não em outra. Um exemplo é a proposição "de" para o português e "*from*" para o inglês. Esses casos eram resolvidos por consenso entre os lexicógrafos em decidir o que é e não é *"stop (sub)word*". Caso semelhante também ocorreu com advérbios como, por exemplo, "como" $^{PT}$, *"how"* $^{EN}$, *"wie"* $^{GER}$.

Sentidos representados por MIDs diferentes é um caso que não causa muito prejuízo ao sistema – e geralmente não continha lexemas de todas as línguas. Na maioria dos casos, notava-se que o indexador "adotava" uma classe de equivalência para a geração da MID. Esse tipo de problema era resolvido simplesmente juntando as classes de equivalência e era comum no começo da construção do tesauro.

MIDs com lexemas de sentidos diferentes normalmente eram gerados pela fusão de classes de equivalências com foco no sentido válido de um lexema para uma língua mas não para outra, de modo a provocar inconsistência na classe de equivalência e até mesmo nos relacionamentos semânticos para outros idiomas – problema típico gerado quando se considera somente os lexemas da língua nativa. Esses problemas eram resolvidos rearranjando os lexemas em outras classes e redefinindo seus relacionamentos semânticos. Por exemplo, *"(heart) murmur"* em diferentes línguas (inglês, alemão e francês) não possui o mesmo

sentido se traduzido de uma língua para outra (*"murmur"* $^{eng}$*, "geraeusch"*$^{ger}$*, ou "souffle"* $^{fra}$ não é a mesma coisa).

Os problemas explicados podem ser considerados como os mais relevantes. Porém, outros tipos de problemas foram encontrados, mas com uma baixa freqüência. Por exemplo, problema de delimitação de *strings* não causa impacto na distribuição das MIDs, e isso se deve ao fato de geralmente tratar de problema relacionado a um lexema de uma língua. Outro caso interessante constitui-se dos casos relacionados à dificuldade de traduzir alguma expressão de uma língua para a outra (e.g., a MID zuiizwjy – *"ZU"* é específico da gramática alemã como *"zurückzukommen"* e não tem tradução para o português) ou a falta de algum lexema referente a uma variante morfológica de um termo já contemplado no tesauro.

Algumas vezes eram encontrados casos em que havia a necessidade de redefinir classes de equivalência que pareciam ter uma espécie de sobreposição conceitual, os quais eram resolvidos por consenso entre os lexicógrafos.

Decisões acerca da resolução de delimitações semânticas, granularidades e sobreposição de conceitos, em alguns casos, representam trabalho intelectual e complexo. Em alguns casos, não compensava incluir todas as acepções para uma determinada classe de equivalência em detrimento do prejuízo muito maior no final da segmentação – isso normalmente era feito com termos exóticos ou fora de contexto da medicina. Um exemplo de difícil solução refere-se à MID "sensiipxrwa" que possui vários sentidos. No dicionário *Yourdictionary*[17] encontram-se as seguintes definições para o português: (1) direção e, (2) sentido. Mas nas outras línguas também encontram-se: inglês: (1) *"intellect"*, "que faz sentido" (2) capacidade do corpo reconhecer as várias sensações (ouvir, sentir, ver, etc), e (3) "*feeling*" (relacionado ao sentido de segurança, alerta). Termos muito ambíguos quando mal resolvido em termo de relacionamento semântico no tesauro causam forte impacto ao SRI, pois o segmentador gera MIDs erradas prejudicando o processo de recuperação de documentos relevantes.

Vale ressaltar que na maioria das vezes, a correção de uma classe, de um relacionamento ou até mesmo um lexema, acabava por levar à correção de outras situações, tornando o gerenciamento dos registros por meio de um protocolo confuso. Uma vez terminadas as correções, foram levantadas as curvas de precisão e revocação. Para a realização dos procedimentos desde a normalização até a geração de dados numéricos para a plotagem dos pontos da curva de precisão e revocação, foram utilizadas ferramentas

[17] *www.yourdictionary.com*

específicas desenvolvidas para tal. A tabela 7 representa um relatório padrão de saída de dados gerados para a montagem da curva de precisão e revocação para todas as versões do tesauro. Essa tabela, da versão de tesauro do dia 23 de agosto de 2005, apresenta a média (*average*) dos valores de precisão para 11 pontos de revocação (0.0, 0.1, 0.2, ..., 1.0). No final da tabela, a média dos 11 pontos de revocação (*11pt average*), dos 3 pontos de revocação (*3pt average*), dos 2 primeiros pontos (*top 2 average*) e dos 3 primeiros pontos (*top 3 average*) para todas as 106 *queries* para essa versão de léxico.

No apêndice são apresentados os resultados do experimento com base na coleção de teste OHSUMED e o referido gráfico[18] com base na versão de tesauro, incluindo o léxico das línguas inglesa, alemã, portuguesa, espanhola (e o sueco), de 23 de agosto de 2005.

Tabela 7: Exemplo de resultados para a versão de tesauro português de 23/08/2005.

QUERIES: dprel_judge_en
MODE tested: 20050823_results_pt

| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| query 1: | 0.0000 | 0.7143 | 0.8333 | 0.5172 | 0.4651 | 0.2525 | 0.2459 | 0.2188 | 0.0000 | 0.0000 | 0.0000 |
| query 2: | 0.0000 | 0.6250 | 0.6250 | 0.4167 | 0.4000 | 0.3125 | 0.3158 | 0.2500 | 0.0000 | 0.0000 | 0.0000 |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| q 104: | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| q 105: | 0.0000 | 0.2300 | 0.2588 | 0.2393 | 0.2028 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| q 106: | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| **Avg:** | **0.4095** | **0.3603** | **0.2775** | **0.2088** | **0.1696** | **0.1413** | **0.0962** | **0.0699** | **0.0338** | **0.0066** | **0.0000** |

11pt average: 0.1612
3pt average: 0.1723
top 2 average: 0.3849
top 3 average: 0.3491

Como se nota na tabela 7, para cada *query* da coleção de teste OHSUMED, processa-se para cada valor de revocação, um valor de precisão; e, ao final desta, calcula-se a média dos valores de precisão para cada ponto de revocação (Avg). Com base nesses valores, são calculados, então, o *11pt average*, que envolve todos os pontos de revocação, o *3pt average*, que envolve três pontos de revocação intermediários, o *top 2 average* que representa a média dos 2 primeiros pontos de revocação e, finalmente, o *top 3 average* que representa a média dos 3 primeiros pontos de revocação. Na figura 23, o eixo da ordenada representa os valores de precisão e o da abscissa os valores de revocação.

[18] Os resultados calculados podem ser verificados na URL http://www.ler.pucpr.br/~roose/dissertation/

Para os propósitos desta dissertação, optou-se em plotar a evolução dos AvgP11 no período considerado do experimento, para uma melhor análise do desempenho da proposta. O figura 23 apresenta o gráfico correspondente à evolução das médicas dos AvgP11 listados na tabela 8.

Tabela 8: Evolução das médias dos valores de precisão sobre 11 pontos de revocação para cada versão de tesauro com base na coleção de teste OHSUMED.

| Pontos de Revocação | Tesauro (aa/mm/dd) | Inglês | Alemão | Português | Espanhol | Sueco |
|---|---|---|---|---|---|---|
| 0.0 | 2005/08/02 | 0.2221 | 0.1905 | 0.1636 | 0.0424 | 0.0297 |
| 0.1 | 2005/08/09 | 0.2211 | 0.1895 | 0.1608 | 0.0408 | 0.0292 |
| 0.2 | 2005/08/16 | 0.2198 | 0.1892 | 0.1601 | 0.0393 | 0.0284 |
| 0.3 | 2005/08/23 | 0.2199 | 0.1885 | 0.1612 | 0.0394 | 0.0333 |
| 0.4 | 2005/09/01 | 0.2204 | 0.1935 | 0.1670 | 0.0561 | 0.0354 |
| 0.5 | 2005/09/08 | 0.2189 | 0.1929 | 0.1669 | 0.0638 | 0.0343 |
| 0.6 | 2005/09/15 | 0.2190 | 0.1932 | 0.1665 | 0.0666 | 0.0358 |
| 0.7 | 2005/09/22 | 0.2192 | 0.1934 | 0.1676 | 0.0661 | 0.0358 |
| 0.8 | 2005/10/01 | 0.2188 | 0.1925 | 0.1667 | 0.0646 | 0.0388 |
| 0.9 | 2005/10/08 | 0.2179 | 0.1955 | 0.1666 | 0.0652 | 0.0390 |

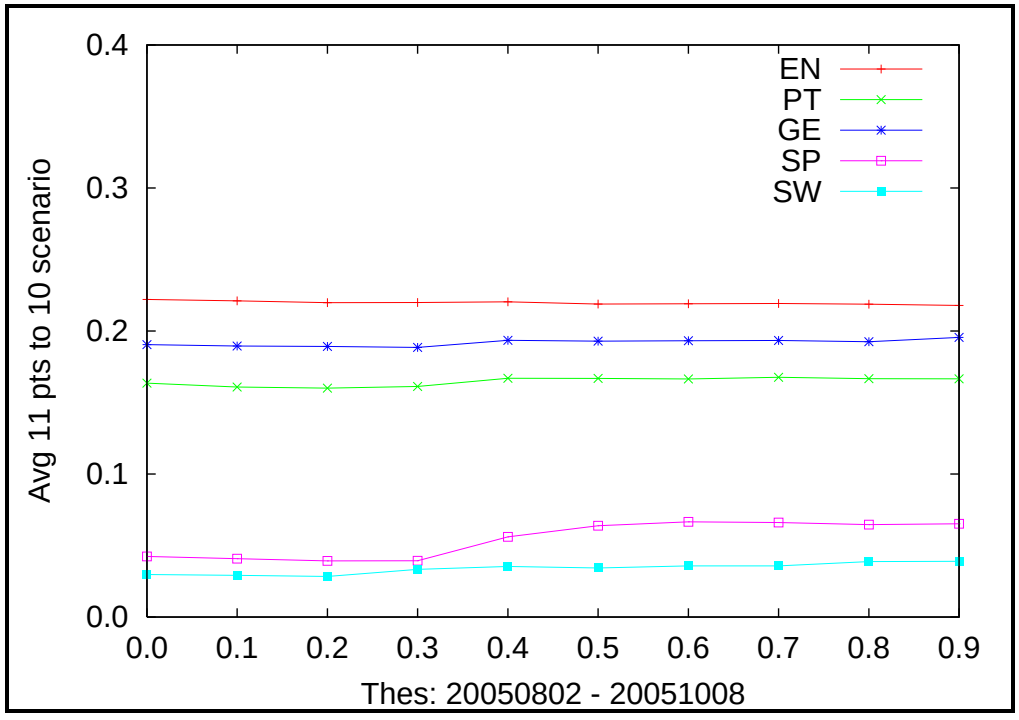



Figura 23: Evolução dos AvgP11 para o léxico inglês, português, alemão, espanhol e o sueco.

# CAPÍTULO 5

# DISCUSSÃO E CONCLUSÃO

## 5.1 DISCUSSÃO

No começo da construção do tesauro, a preocupação inicial era focada basicamente na quantidade de entradas, ou seja, criação de classes de equivalências e seu incremento com *subwords* sinônimas. Numa segunda etapa, o foco voltou-se ao melhoramento, no sentido de corrigir segmentações errôneas normalmente ligadas às questões sintáticas ou à criação de novas classes de equivalências. Na terceira etapa, com o léxico possuindo uma boa cobertura da terminologia médica, as atividades foram direcionadas à realização das tarefas para caracterizar, de fato, um tesauro; ou seja, configurar os diversos tipos de relacionamentos entre as classes de equivalências quando necessárias, conforme explicado no ítem 2.7.

Apesar de haver uma boa comunicação entre os lexicógrafos e, em alguns casos, existir aprovação unânime para estabelecer novas relações semânticas ou realizar alguma modificação, não era suficiente para evitar problemas de explosão de relacionamentos semânticos encadeados (*chains*). Outro tipo de problema comum envolvia relacionamento entre classes de equivalência que fechavam um "ciclo" (*cycle*); ou seja, pelo fato de não se ter condições visuais a todos os relacionamentos entre classes de equivalências, era comum a formação de células circulares. Esses dois problemas causavam sérios problemas para outro módulo desambiguador do *Morphosaurus*. De um modo geral, quando se descobria algum tipo de problema no tesauro através de um caso, procurava-se, então, levantar os casos semelhantes para resolvê-los. O problema é que normalmente os outros tipos de problemas ficavam mascarados. Mediante esses fatos, sentiu-se a necessidade de um método que apontasse de forma mais sistemática qualquer tipo de problema no resultado final, na geração da representação dos significados dos grupos de lexemas, isto é, das MIDs. Dessa forma, com objetivo de incrementar a qualidade do tesauro, decidiu-se utilizar *corpora* comparáveis como ponto de partida para a detecção classes de equivalência com potenciais problemas – nesse processo algumas classes não apresentaram problemas. Como exemplo, pode-se citar a MID physioterapriirzja (krankengymnast$^{GER}$, physiotherap$^{EN}$ fisioterap$^{PT}$ fysioterap$^{SW}$). Esse tipo de fato representou 10% dos casos para o português e o alemão – um índice relativamente alto se comparado com outros tipos de problemas, conforme mostrado na tabela 6. Uma explicação

plausível para esse fenômeno decorre dos termos ambíguos que possuem uma ocorrência maior numa língua que na outra. O segmentador retorna os sentidos normalizados de um termo ambíguo, por exemplo, para o termo "lobo$^{PT}$", o segmentador retorna as MIDS "lobiikiwqa e wolfijyjkpa". Num processo estatístico, pode-se tomar duas estratégias para a geração das listas de freqüências: (a) aplicar alguma técnica para resolver a ambigüidade do resultado e utilizar o termo correto nos cálculos estatísticos ou, (b) utilizar todos os termos normalizados nos cálculos estatísticos. O ideal seria dispor de um desambiguador de forma a aproximar a análise do termo ambíguo o mais próximo possível do contexto na qual estaria inserido, porém, haja vista a dificuldade na implementação de tais ferramentas, optou-se por implementar um desambiguador simples no qual foram contadas as ocorrências mais comuns e as freqüências mais altas foram utilizadas como fator determinante para resolver os casos ambíguos. Isso explicaria o motivo do porquê existirem algumas MIDs no topo da lista de freqüência sem apresentar problemas.

No começo dos experimentos, havia a expectativa de haver incrementos significativos após as correções das MIDs seguindo a lista proposta na metodologia. Apesar disso, os resultados mostraram incrementos muito pequenos no que diz respeito ao parâmetro precisão. Comparando os primeiros valores de AvgP11 com as últimas calculadas no processo, o crescimento é relativamente insignificante para o português e o alemão, com valores de 1.8% e 2.6%, respectivamente. Aparentemente, esse pequeno incremento parece estar relacionado, principalmente, à criação de novos relacionamentos semânticos e alguns rearranjos, uma vez que o léxico destes pode ser considerado consolidado, ou seja, um léxico com boa cobertura do domínio médico. Por outro lado, o desempenho de RI com o idioma inglês teve um decremento de 1.9%. Esse valor pode ser considerado como um valor normal dentro de uma tolerância de variação, assim como ocorreu com os idiomas português e alemão, ainda mais se for considerado que o *benchmark* montado não mede todo o universo da informação, mas o desempenho da RI de uma amostra de 106 *queries*.

Certamente, quanto mais consolidado um tesauro, menor o impacto no desempenho da RI a uma modificação no léxico. Por outro lado, o incremento no desempenho do *benchmark* espanhol alcançou 53% com relação ao seu valor inicial de Avg11, e não se pode creditar às operações de relacionamentos semânticos. Esse desempenho leva a interpretar que essa metodologia é adequada para a escolha de casos mais graves de representações semânticas a serem corrigidos, surtindo também melhora na produtividade das correções.

## 5.2 CONCLUSÕES

Nessa dissertação, desenvolveu-se uma metodologia que auxilia a manutenção de um tesauro multilíngüe para a área médica, por meio de amostra representativa de textos bilíngües comparáveis para a detecção de potenciais representações ou classes de sinônimos ou relacionamentos semânticos que venham a prejudicar o desempenho do processo de recuperação de documentos médicos relevantes. A técnica pode ser aplicada com a utilização de *corpora* comparáveis e apresentou progressos na qualidade do tesauro utilizando um benchmark de RI.

Implementar um sistema de recuperação de informações é de fato um trabalho demorado, caro e complexo se for baseado num tesauro. Além do mais, em se tratando de um sistema multilíngüe, é necessário um ambiente multidisciplinar onde os integrantes estejam de fato comprometidos com a qualidade do mesmo. É um trabalho que não tolera erros graves de relacionamentos no tesauro e nem de faltas graves no léxico, sob pena de resultados desastrosos no desempenho da máquina de busca e mau desempenho no sistema de recuperação como um todo.

Lidar com a representação de sentidos de expressões lingüísticas através de representações simbólicas padronizadas é complicado conforme exposto nesse trabalho. Especificamente com a representação textual, a dificuldade mantém-se pela natureza diversificada, advindo de fenômenos lingüísticos e da dependência contextual.

Enquanto a linguagem natural é extremamente fácil para seres humanos, o entendimento dela por sistemas de computadores, mesmo com a aplicação de técnicas de PLN aliadas com alguma abordagem de RI, é uma tarefa árdua no campo da computação. A linguagem natural permite uma variedade de subterfúgios que as técnicas computacionais ainda não conseguem cobrir. Com base nas 160 MIDs resolvidas, pode-se resumir os problemas típicos nos seguintes casos: (a) mesmo conceito expresso de formas diferentes, (b) mesma representação simbólica que pode ter diferentes significados, (c) ambigüidade de interpretação de um símbolo, (d) mesmo conceito que pode ter diferente significado dependendo do contexto e (e) expressões vagas, desprovidas de especificidade.

Lidar com o processamento da linguagem natural requer diferentes tipos de conhecimentos e o processamento computacional que extrai e processa texto não lidam com o entendimento. Apesar da área de PLN utilizar recursos estatísticos e matemáticos, existem ainda muitos desafios a serem resolvidos quando se trata de abordagens simbólicas para o

processamento da linguagem natural. O problema continua sendo a complexidade da representação do conhecimento.

A proposta aqui mostrada é um pequeno passo na solução de um problema pontual – e, mesmo assim, não resolve tudo, pois ele é sistemático na verificação de um resultado final, que é a representação por uma língua artificial. Os incrementos não foram significativos, pois trabalhou-se na "curva de saturação", ou seja, com um tesauro consolidado e tempo de acompanhamento limitado. Mas, pode-se constatar uma melhora significativa relativa das línguas espanhola e sueca, que estão em fase de construção, com relação aos léxicos consolidados: o inglês, o português e o alemão. De qualquer forma, os procedimentos propostos pela abordagem servem para balizar a construção de um tesauro com um mínimo de erros e, assim, almejar um padrão de qualidade que se reflita na recuperação efetiva de documentos relevantes.

Além de ajudar na monitoração da construção e manutenção do tesauro, os procedimentos também reduziram o tempo despendido na detecção dos erros que anteriormente se realizava de forma visual através dos resultados da segmentação de listas de termos médicos compilados.

A metodologia de confrontar amostras de textos normalizados pelo sistema *Morphosaurus*, que pode ser estendida a *corpora* comparáveis, mostrou-se efetiva para expor de forma direta os problemas contemplados no tesauro.

Pelo fato de se tratar com questões subjetivas, como é o caso de resolver ambigüidades e outros aspectos oriundos de fenômenos lingüísticos, esta metodologia constitui-se numa ferramenta para amenizar o processo do gerenciamento do tesauro no que diz respeito à sua monitorização, resultando na diminuição do ruído no sistema.

Neste trabalho, utilizou-se o Sistema *Morphosaurus* como *workbench*, mas poderia ser qualquer outro que empregue um tesauro, mesmo sob outro enfoque, para o mapeamento de documentos multilíngüe na representação artificial empregando descritores semânticos.

O processo de construção de um tesauro, assim como qualquer processo de construção, envolve controle de qualidade. Neste trabalho destacou-se, entre outras coisas, a dificuldade na montagem de um tesauro e a necessidade de uma metodologia que mantenha o seu gerenciamento de forma a minimizar os erros. Em primeiro lugar, a proposta explicita a grande maioria dos descritores com problemas reais a serem corrigidos pelos lexicógrafos, refletindo na produtividade da manutenção. Assim, conclui-se que a metodologia integrada a um *workflow* na manutenção de um tesauro reflete também na qualidade de um Sistema de Recuperação de Informações.

## 5.3 TRABALHOS FUTUROS

A grande maioria dos trabalhos sobre avaliação de SRI está relacionada ao desempenho de um sistema de recuperação de informações como um todo. Pesquisas que enfocam qualidade de tesauro são raras. Sugere-se, então, dar continuidade ao refinamento da qualidade do tesauro, englobando os idiomas espanhol, sueco e francês, de forma a equalizar suas coberturas lexicais ao nível das línguas inglesa, alemã e portuguesa.

Atualmente, desenvolve-se a implementação de ferramentas de linguagem natural como os etiquetadores (*taggers*). Independente da representação adotada, sempre haverá fenômenos lingüísticos como uma barreira a ser vencida e, desta forma, a área de PLN apresenta-se como mais um aliado para a melhora da busca de documentos relevantes. Certamente, a utilização de etiquetadores no sistema *Morphosaurus* incrementará a qualidade da busca de documentos. A aplicação da metodologia desenvolvida nesse trabalho faz-se efetiva para mensurar a qualidade de um tesauro multilíngüe. Nesta dissertação, utilizou-se 106 *queries* para a avaliação do desempenho, e isso não é suficiente para medir todo o universo da informação, mesmo *in loco*. A linguagem é probabilística e os meios de avaliação são subjetivos. Assim, para alcançar índices que expressem a realidade, sugere-se técnicas de avaliação que englobem tanto os documentos quanto as *queries*.

# APÊNDICE

## CURVA DE PRECISÃO E REVOCAÇÃO PARA O TESAURO DE 23/08/2005

As tabelas abaixo apresentam os resultados dos cálculos gerados no processamento da coleção de teste OHSUMED normalizadas nas línguas inglesa, alemã, portuguesa, espanhola e sueca com base nas versões de tesauro de 23/08/2005. As explicações sobre elas podem ser vistas no item 4.1.

Tabela 9: Resultados para o tesauro de 23/08/2005 para as *queries* inglesa

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QUERIES: | dprel_judge_en | | | | | | | | | | |
| MODE tested: | 20050823_results_en | | | | | | | | | | |
| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| avg: | 0.5048 | 0.4494 | 0.3661 | 0.3028 | 0.2433 | 0.2081 | 0.1510 | 0.1127 | 0.0578 | 0.0165 | 0.0060 |
| 11pt average: | 0.2199 | | | | | | | | | | |
| 3pt average: | 0.2407 | | | | | | | | | | |
| top 2 average: | 0.4771 | | | | | | | | | | |
| top 3 average: | 0.4401 | | | | | | | | | | |

Tabela 10: Resultados para o tesauro de 23/08/2005 para as *queries* alemã

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QUERIES: | dprel_judge_en | | | | | | | | | | |
| MODE tested: | 20050823_results_ge | | | | | | | | | | |
| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| avg: | 0.4000 | 0.3791 | 0.3114 | 0.2535 | 0.2054 | 0.1800 | 0.1376 | 0.1045 | 0.0647 | 0.0313 | 0.0060 |
| 11pt average: | 0.1885 | | | | | | | | | | |
| 3pt average: | 0.2071 | | | | | | | | | | |
| top 2 average: | 0.3896 | | | | | | | | | | |
| top 3 average: | 0.3635 | | | | | | | | | | |

Tabela 11: Resultados para o tesauro de 23/08/2005 para as *queries* portuguesa

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QUERIES: | dprel_judge_en | | | | | | | | | | |
| MODE tested: | 20050823_results_pt | | | | | | | | | | |
| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | |
| avg: | 0.4095 | 0.3603 | 0.2775 | 0.2088 | 0.1696 | 0.1413 | 0.0962 | 0.0699 | 0.0338 | 0.0066 | 0.0000 |
| 11pt average: | 0.1612 | | | | | | | | | | |
| 3pt average: | 0.1723 | | | | | | | | | | |
| top 2 average: | 0.3849 | | | | | | | | | | |
| top 3 average: | 0.3491 | | | | | | | | | | |

Tabela 12: Resultados para o tesauro de 23/08/2005 para as *queries* espanhola

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| QUERIES: | dprel_judge_en | | | | | | | | | |
| MODE tested: | 20050823_results_sp | | | | | | | | | |
| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| avg: | 0.0962 | 0.0875 | 0.0654 | 0.0495 | 0.0414 | 0.0352 | 0.0254 | 0.0218 | 0.0098 | 0.0014 | 0.0000 |
| 11pt average: | 0.0394 | | | | | | | | | |
| 3pt average: | 0.0429 | | | | | | | | | |
| top 2 average: | 0.0918 | | | | | | | | | |
| top 3 average: | 0.0830 | | | | | | | | | |

Tabela 13: Resultados para o tesauro de 23/08/2005 para as *queries* sueca

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| QUERIES: | dprel_judge_en | | | | | | | | | |
| MODE tested: | 20050823_results_sw | | | | | | | | | |
| Recall: | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| avg: | 0.0667 | 0.0871 | 0.0658 | 0.0392 | 0.0316 | 0.0272 | 0.0224 | 0.0160 | 0.0071 | 0.0028 | 0.0000 |
| 11pt average: | 0.0333 | | | | | | | | | |
| 3pt average: | 0.0378 | | | | | | | | | |
| top 2 average: | 0.0769 | | | | | | | | | |
| top 3 average: | 0.0732 | | | | | | | | | |

Em seguida, apresenta-se o gráfico de precisão e revocação das versões de léxico de 23/08/2005 nas línguas inglesa, alemã, portuguesa, espanhola e sueca.

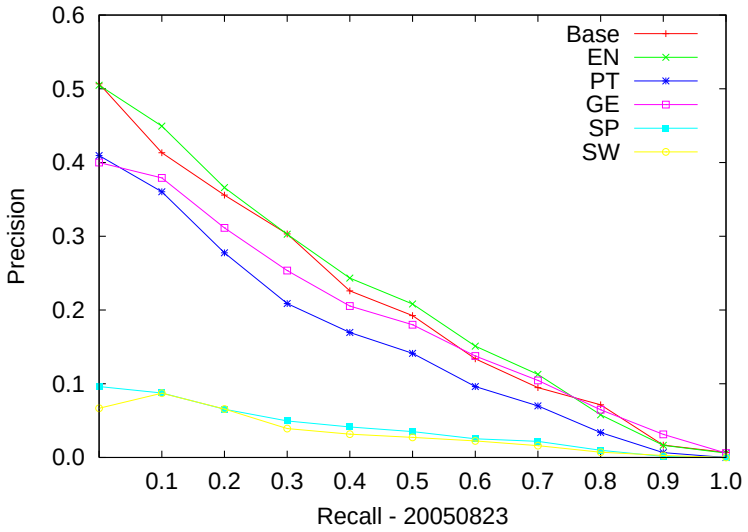



Figura 24: Gráfico de precisão e revocação para a versão de léxico de 23/08/2005 para as línguas inglesa, portuguesa, alemã, espanhola e sueca.

## REFERÊNCIAS BIBLIOGRÁFICAS


ABAITUA, J. Tratamiento de corpora bilingües. In: M. A. Martin (Eds.). **Tratamiento del lenguaje natural**. Barcelona: Univesitat de Barcelona, p. 61-90, 2002.

ABEL, M. Estudo da Perícia em petrografia sedimentar e sua importância para a engenharia de conhecimento. (**Tese de Doutorado**). Programa de Pós-Graduação em Computação, UFRGS, Porto Alegre, 2001.

AIRES, R. **Avaliação em Recuperação de Informação**. Portugal, 2002.

ANDRADE, R. L., G. N. NOGUEIRA-NETO*, et al.* Recuperação Translingual de Textos via Representação Interlingual. **Congresso Brasileiro de informática em Saúde**. Ribeirão Preto, São Paulo: Sociedade Brasileira de Informática em Saúde, v. 1, p. 1202-1207, 2004.

ATKINS, J. C. e N. OSTLER. **Corpus Design Criteria**. Oxford: Oxford University Press, 1992.

BAEZA-YATES, R. e B. RIBEIRO-NETO. **Modern Information Retrieval**. New York: Addison Wesley Longman Publishing Co, 1999.

BAR-HILLEL, Y. e R. CARNAP. Semantic Information. **Philo Sci**, v. 4, p. 147-157, 1953.

BELKIN, N. J. e W. B. CROFT. Information Filtering and Information Retrieval: Two Sides of the Same Coin? Comunication of the ACM, v. 35, n. 12, p. 29-38, 1992.

BHOLA, H. S. **Evaluating "Literacy for development" projects, programs and campaigns: Evaluation planning, design and implementation, and utilization of evaluation results**. Hamburg, Germany: UNESCO Institute for Education; DSE (German Foundation for International Developement), 1990.

**BLIKSTEIN, I. Kaspar Hauser ou a fabricação da realidade**. São Paulo: Cultrix, 1990.

BUCKLAND, M. **Information and Information System**. New York: Greenwood, 1991.

CARVALHO, E. C. A natureza social da Ciência da Informação. In: L. V. R. Pinheiro (Eds.). **Ciência da Informação, Ciências Sociais e Interdisciplinaridade**. Rio de Janeiro: IBICT, p. 51-53, 1999.

CINTRA, A. M. M. **Para entender as linguagens documentárias**. São Paulo: Polis, 2002.

DÉJEAN, H., E. GAUSSIER*, et al.* An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. **Proceedings of the 19th international conference on Computational linguistics**. Taipei, Taiwan: Association for Computational Linguistics, p. 1-7, 2002.

ECO, U. **Semiótica e filosofia da linguagem**. Editora Ática, 1996.

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. Cambridge, MA: MIT Press, 1998.

FOSKETT, D. J. Thesaurus. In: D. J. Foskett (Eds.). **Reading in Information Retrieval**. New York: Morgan Kaufmann, p. 111-134, 1997.

FRIEDMAN, C. e G. HRIPCSAK. Natural language processing and its future in medicine. **Acad Med**, v. 74, n. 8, Aug, p. 890-5. 1999.

FUHR, N. Probabilistic Models in Information Retrieval. **Computer Jornal**, v. 35, n. 3, p. 243-255. 1992.

FUNG, P. A statistical view of bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: J. Véronis (Eds.). **Parallel Text Processing**. 2000.

FURNAS, G. W. E. A. The vocabulary problem in human-system communication. **Comunications of the ACM**, v. 11, 1987.

GALLE, M., O. JAKOBS*, et al.* Dokumentation des studienprojektes - aufbereitung des dpa korpus: University of Trier, 1992.

GOVE, P. B. **Webster's Third New International Dictionary**. Springfield, MA: Merriam-Webster Inc., 1986.

GREFENSTETTE, G. Cross-Language Information Retrieval. In: W. B. Croft (Eds.). **The Kluwer International Series on Information Retrieval**. Grenoble, France, 182 p., 1998.

GWIZDKA, J. e M. CHIGNELL. **Towards information Retrieval Measure for Evaluation of Web Search Engines**. 1999.

HAHN, U., S. SCHULZ*, et al.* Crossing Languages in Text Retrieval via an Interlingua. **Recherche d'Information Assistée par Ordinateur - RIAO 2004**. Avignon l'Université d'Avignon, p. 100-115, 2004.

HALLIDAY, M. A. K. Corpus Studies and Probabilistic Grammar. **AIJMER, K.; Altenberg, B. (orgs.). English Corpus Linguistics: Studies in honour of Svartvik**. Londres: Longman, p. 30-43, 1991.

HAYAKAWA, S. I. **Language in Thought and Action**. New York: Harcourt, Brace & World, 1939.

HEARST, M. A. The Use of Categories and Clusters for Organizing Retrieval Results. In: T. Strzalkowski (Eds.). **Natural Language Information Retrieval**. Dordrecht: Kluwer Academic Publishers, v.7, p. 333-374, 1999.

HEATON, J. **Programming Spiders, Bots, and Aggregators in Java**. San Francisco: Sybex, 2002.

HECKERLING, P. S. Information Content of Diagnostic Tests in the Medical Literature **Methods Inf. Med.**: Pubmed- Medline, v. 29, p. 61-66, 1990.

HERSH, W. R. **Information Retrieval - A Health Care Perspective**. New York: Springer, 1996.

HERSH, W. R., C. BUCKLEY*, et al.* OHSUMED: An interactive retrieval evaluation and new large test collection for research. **Proceedings of the 17th Annual ACM SIGIR Conference**, p. 192-201, 1994.

HUGE, G. Combining Corpus Linguistics and Human Memory models for Automatic Term Association. In: T. Strzalkowski (Eds.). **Natural Language Information Retrieval**, p. 75-98, 1999.

JESUS, J. B. M. D. Tesauro: Um Instrumento de Representação do Conhecimento em Sistemas de Recuperação do Conhecimento em Sistemas de Recuperação de Informação. **Anais do XII Seminário Nacional de Bibliotecas Universitárias**. Recife: Universidade Federal de Pernambuco. 2002.

LIMA, V. M. A. Terminologia, Comunicação e Representação Documentária. (**Mestrado**). Escola de Comunicação e Artes (ECA), Universidade de São Paulo - USP, São Paulo, 1998.

LOSEE, R. M. **The Science of Information: Measure and Applications**. San Diego, CA: Academic Press, 1990.

MANNING, C. D. e H. SCHÜTZE. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MARCHIONINI, G. Interface for end-user information seeking. **J Am Soc Info Sci**, n. 43, p. 156-163. 1992.

MASON, O. **Programming for Corpus Linguistics - How to Do Text Analysis with Java**. Edinburgh: Edinburgh University Press, 2000.

MEADOW, C. T., B. R. BOYCE*, et al.* **Text Information Retrieval System**. Los Angeles: Academic Press, 1992.

MILLER, U. Thesaurus construction: problems and their roots. **Information Processing & Management**, v. 33, p. 481-493, 1997.

MIRANDA, A. Globalización y sistemas de información: nuevos paradigmas y nuevos desafios. Disponível em: http://eprints.rclis.org/archive/00003663/. Acessado em 11/01/2006.

MIZZARO, S. A Cognitive Analysis of Information Retrieval. **Information Science: Integration in Perspective, CoLis2**: The Royal School of Librarianship, p. 233-250, 1996.

OARD, D. W. Alternative approaches for cross-language text retrieval. **Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval**, 1997.

OGATA, K. **Engenharia de Controle Moderno**. Rio de Janeiro: Prentice Hall do Brasil, 1990.

OGDEN, C. K. e I. A. RICHARDS. **The Meaning of Meaning**. New York: Hartcourt, Brace & Co., 1956.

OLIVEIRA, D. H. **Introdução a XML e suas aplicações**. 2002.

PERCY, C. E. e C. F. MEYER. Synchronic Corpus Linguistics. **papers from the sixteenth International Conference on English Language and Research on Computerized Corpora (ICAME 16)**. Amsterdã, 1996.

PETERS, C. Cross-language Information Retrieval - Revised papers of the Workshop of the Cross-language Information Retrieval. **LNCS 2069, Forum CLEF**. Lisboa, Portugal, 2000.

RAPP, R. Identifying word translations in nonparallel texts. **Proceedings of the Annual Meeting of the ACL**, 1995.

RIJSBERGEN, C. J. V. **Information Retrieval**. London: Butterworth, 1979.

RIJSBERGEN, C. J. V., M. LALMAS*, et al.* Information Retrieval and Situation Theory. **ACM SIGIR Forum**. New York, v. 30, p. 11-25, 1996.

SABATER, J. e C. SIERRA. Review on Computational Trust and Reputation Models. **Artificial Intelligence Review**, v. 24, n. 1, p. 33-60. 2005.

SALTON, G. **The SMART Retrieval System**. Englewood Clifs, N.J.: Prentice Hall, Inc., 1971.

SALTON, G., MACGILL, M. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SANCHES, A., CANTOS P. Predictability of Word forms (types) and Lemmas in Linguistic Corpora. A case study based on analysis of the COMBRE Corpus: an 8 -million word corpus of contemporary Spanish. **International Journal of Corpus Linguistics**, Amsterdã, p. 258-280. dez/1997.

SARDINHA, T. B. **Lingüística de Corpus**. Tamboré, SP: Manole, 2004.

SCHULZ, S. e U. HAHN. Morpheme-based cross-language indexing for medical document retrieval. **International Journal of Medical Informatics (IJMI)**, v. 58, n. 59, p. 87-99. 2000.

SCHULZ, S., HAHN, U. Syntatic and Semantic Aspects of Subword Indexing. **International Journal of Medical Informatics (IJMI)**. Italy, 2006.

SCHULZ, S., K. MARKÓ*, et al.* Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. **COLING Geneva 2004 -**

**Proceeding of the 20th International Conference on Computational Linguistics**. Switzerland: Association for Computational Linguistics, v. 2, p. 813-819, 2004.

SEATON, A. F. Low level Language Processing for Large Scale Information Retrieval: What techniques actually work. **In Procceding of Workshop on Terminology, Information Retrieval and Linguistics**. Rome, Italy, p. 69-77, 1995.

SHANNON, C. E. e W. WEAVER. **The Mathematical Theory of Communication**. Urbana: University of Illinois Press, 1949.

SINCLAIR, M. From Theory to Practice. Spoken english on computer: transcription, mark-up and applicaton. In: M. G. Leech G., Thomas J. (Eds.). Londres: Logman, 1995.

SOERGEL, D. Functions of a thesaurus - classification, ontological knowledge base: College of Library and Information Services. University of Maryland, 1997.

STRZALKOWSKI, T. **Natural Language Information Retrieval**. Kluwer Academic Publishers, 1999.

TARDELLI, A. O., M. S. ANCAO*, et al.* Descoberta baseada em literatura: Um enfoque experimental para descoberta aberta em bases de dados do tipo MEDLINE. **VIII Congresso Brasileiro de Informática em Saúde - CBIS 2002**. Natal - RN: SBIS, 2002.

UMLS. Knowledge Sources. **Unified Medical Language System**: Unified Medical Language System - U.S. Departament of Health and Human Services, National Institutes of Health, National Library of Medicine, 1994.

UMLS. Bethesda, MD. **National Library of Medicine, Unified Medical Language System**, 2005

WITTEN, I. H., MOFFAT, A., BELL, T. **Managing gigabytes: compressing and indexing documents and images** New York: Van Nostrand Reinhold, 1994.

WIVES, L. K. Utilizando conceitos como descritores de textos para o processamento de identificação de conglomerados (clustering) de documentos. (**Tese de Doutorado**). Programa de Pós-Graduação em Computação, UFRGS, Porto Alegre, 2004.

YERGEAU, F., G. ADAMS*, et al.* Internationalization of the Hypertext Markup Language. RFC 2070: Network Working Group, 1997.

ZHANG, D., N. K. RODERER*, et al.* Developing a UMLS-based Indexing Tool for Health Science Repository System. **AMIA Annu Symp Proc**, p. 1157. 2006.

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENG[a] ELÉTRICA E INFORMÁTICA INDUSTRIAL

# "Detecção de Erros em Tesauro Médico Multilíngüe Através de Corpora Comparáveis"

por

## Roosewelt Leite de Andrade

**Esta Dissertação foi apresentada no dia 20 de Dezembro de 2006, como requisito parcial para a obtenção do grau de MESTRE EM CIÊNCIAS – Área de Concentração: Engenharia Biomédica. Aprovada pela Banca Examinadora composta pelos professores:**

**Prof. Dr. Percy Nohama**
**(Orientador – UTFPR)**

**Prof. Dr. Stefan Schulz**
**(ALUF)**

**Profª. Drª. Andreia Malucelli**
**(PUC-PR)**

**Profª. Drª. Cláudia M. Cabral Moro Barra**
**(PUC-PR)**

**Visto e aprovado para impressão:**

**Prof. Dr. Hugo Reuters Schelin**
**(Coordenador do CPGEI)**

**RESUMO:**

A terminologia médica é complexa e esse fenômeno exerce um impacto forte na construção e manutenção de um tesauro do domínio médico. Metodologias para o controle de qualidade são de extrema importância, pois permitem detectar erros e consequentemente melhorar o desempenho de aplicações que utilizam tesauros, como, por exemplo, os Sistemas de Recuperação de Informações. Neste trabalho, propõe-se uma nova metodologia para a monitoração da construção e manutenção de um tesauro médico multilíngüe baseado em *subwords* através da utilização de *corpora* comparáveis para a detecção de descritores semânticos com problemas. Isso foi realizado comparando o perfil de distribuição de freqüência, em pares, dos descritores de um tesauro e verificaram-se os desequilíbrios na distribuição de ocorrências dos descritores semânticos para os idiomas português-inglês e alemão-inglês para serem corrigidos pelos lexicógrafos. Após as correções, uma avaliação sumativa foi realizada pela medida de parâmetro de desempenho que utiliza um *benchmark* de recuperação de informações padrão. A metodologia identificou problemas típicos como ausência de descritores semânticos, descritores diferentes com mesmo sentido, mesmo descritor com sentidos diferentes e ambigüidade dependente do idioma. Avaliando o desempenho na recuperação de informação, sobre o período do experimento, constatou-se um crescimento relativamente pequeno para os valores de precisão e revocação referente ao português e ao alemão. Houve um pequeno decremento para a língua inglesa, em contraste com o desempenho notável para a língua espanhola que alcançou um índice de 50%, em relação ao estado inicial dos valores de precisão, em três meses. Conclui-se que esse método é efetivo para a identificação de descritores com problemas e recomenda-se sua integração às operações de manutenção de um tesauro.

PALAVRAS-CHAVE

Vocabulário Controlado, Recuperação de Informação Multilíngüe, Controle de Qualidade, Informação em saúde.

ÁREA/SUB-ÁREA DE CONHECIMENTO

1.03.03.04 – 9 Sistemas de Informação

1.03.03.02 – 2 Engenharia de Software

6.07.02.03 – 6 Técnicas de Recuperação de Informação

2006

Nº: 432

# Livros Grátis

( http://www.livrosgratis.com.br )

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo