PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE

JEFERSON LUIZ BITENCOURT

MONITORAÇÃO DE PROCEDIMENTOS EM UM TESAURO MULTILÍNGÜE

DISSERTAÇÃO DE MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO
EM TECNOLOGIA EM SAÚDE

CURITIBA
2006

JEFERSON LUIZ BITENCOURT

MONITORAÇÃO DE PROCEDIMENTOS EM UM TESAURO MULTILÍNGÜE

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do Grau de Mestre em Tecnologia em Saúde. Área de Concentração: Informática em Saúde

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

CURITIBA
2006

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

B624m 2006

Bitencourt, Jeferson Luiz
Monitoração de procedimentos em um tesauro multilíngüe / Jeferson Luiz Bitencourt ; orientador, Percy Nohama ; co-orientador, Stefan Paul Schulz. -- 2006.
82 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2006
Bibliografia: f. 71-77

1. Medicina - Terminologia. 2. Tesauro. 3. Sistema de recuperação da Informação – Medicina. 4. Indexação. I. Nohama, Percy. II. Schulz, Stefan Paul. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Tecnologia em Saúde. IV. Título.

CDD 20. ed. – 025.4961

Dedicatória

Primeiramente, a Deus por ter me ajudado neste momento tão importante e complicado da minha vida, dando-me força e iluminando-me para que eu terminasse este trabalho, carregando-me no colo conforme a sua palavra.

Aos meus pais, Luiz Gonzaga Bitencourt e Lelia de Fátima Arruda Bitencourt, que sempre me apoiaram e acreditam em mim, dando-me uma educação que hoje considero muito importante, sendo os meus dois exemplos de vida, os quais quero sempre seguir.

À minha esposa Nadia Cristina Estella Kaspchak Bitencourt, o amor da minha vida, por toda a força e apoio, que muitas vezes foi a minha base forte para que eu não desistisse.

Aos meus irmãos, Jeison Willian Bitencourt e Cleverson Bitencourt, que me ajudam a cada dia, não só pelos incentivos, mas também por estarmos sempre juntos e formarmos uma família maravilhosa.

Aos meus familiares, por me ajudarem em todos os sentidos, pelas orações apoio e, ainda, por entenderem minha ausência no decorrer do mestrado.

Aos amigos que me ajudaram direta ou indiretamente, tanto no trabalho quanto nas palavras de motivação.

Agradecimentos

Ao Prof. Dr. Percy Nohama, meu agradecimento especial. Mais que um orientador, um exemplo de dedicação, exigência e compreensão. Muito obrigado pelo voto de confiança, por ter acreditado e pelas palavras de incentivos no momento que mais precisei.

Ao Prof. Dr. Stefan Paul Schulz, co-orientador, por sua contribuição e apoio, que foram de extrema importância para conclusão deste trabalho. Gostaria de agradecer também aos pesquisadores da Alemanha, pois sem eles este trabalho não aconteceria.

Aos meus pais, que fizeram de tudo por mim, investindo boa parte de suas vidas, na minha educação e dos meus irmãos. Todo esse sacrifício permitiu-nos ter o que jamais eles puderam ter. Aos meus irmãos, grandes amigos, com os quais posso contar em cada fase da minha vida.

À Profa. Andreia Malucelli, pelo tanto que me ajudou, pela maneira sempre atenciosa, divertida e educada de se relacionar e por estar sempre disposta a compartilhar. Nós sabemos o quão importante é a sua presença na minha educação. Aos professores que, além de simplesmente exporem as matérias em aula, sanarem-me em dúvidas e estiveram sempre abertos a dar qualquer explicação.

À minha esposa, pela paciência, compreensão, incentivo e carinho.

Aos colegas do Laboratório de Engenharia de Reabilitação (LER): Prof. Edson José Pacheco, Adriano Ricardo Duma, Píndaro Secco Cancian, Guilherme Nunes Nogueira Neto, Hood Wilson Gusso da Silva, Michel Oleynik, Luciana Bandeira, Thais Ariela Machado, que me ajudaram a realizar esta pesquisa e, principalmente, ao Roosewelt Leite de Andrade, pelo incentivo, apoio constante e disposição em resolver problemas.

Ao CNPq, pelo apoio para realização deste trabalho.

Enfim, a Deus, aos meus familiares e a todas as pessoas que de alguma forma contribuíram para a realização desta pesquisa.

“Você que conquista novos horizontes
A garimpar sabedoria como fonte
Sempre foi, e é um sonho seu.
Lute com perseverança
Tenha em Deus confiança
Tudo isso é dom de Deus

Não nascemos senão para a luta
Não fique só as escutas
Lutar é nosso dever
Você quer vencer a escalada
Porque descobriu na jornada
Um tesouro dentro de você

Cultive sempre esse tesouro
Que vale mais que o ouro
Antes era escondido
Use com serenidade
Nos campos, nas favelas, nas cidades.
Seja você um comprometido

Seguistes um belo caminho
Com um florescente destino
Não abandones jamais
Abraças-te a sabedoria
Fez explodir alegria
Para o orgulho dos teus pais.

Jeferson, isso que pude te oferecer,
o vô tem uma coisa a dizer,
Volte a teu campo escavar,
Esse tesouro é pra poucos,
Para sábios e para loucos,
Deus vai te abençoar”

Felício José Bittencourt
(2006)

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS

An del – Anomalia de delimitação

An per – Anomalia de permanência

An rel – Anomalia de relacionamento

An tip - Anomalia de tipo

BDA - base de dados antigo

BDN - base de dados novo

Ceq – Classe de equivalência

JDBC - *Java Database Connectivity*

JSP - *Java Server Pages*

LN - Linguagem Natural

MESH - *Medical Subject Headings*

MID - *MorphoSaurus Idenfier*

PLN - processamento da Linguagem Natural

RI – Recuperação de informação

SRI – Sistema de Recuperação de informação

UMLS - *Unified Medical Language System*

W3C - *World Wide Web Consortium*

WWW - *World Wide Web*

XML - *Extensible Markup Language*

RESUMO

Construir um tesauro na área médica com a idéia de reunir classes de sinônimos e contemplar acepções não é uma tarefa trivial, devido à complexidade inerente à própria terminologia. Nesta dissertação, abordam-se problemas existentes na criação manual de um tesauro e descreve-se uma técnica de apoio ao lexicógrafo na inclusão e manutenção de lexemas. Introduz-se o conceito de anomalias com procedimentos realizdos pelos lexicógrafos sem impacto positivo à qualidade do tesauro. Mapearam-se 4 tipos de procedimentos incorretos: anomalias de relacionamento, tipo, delimitação e permanência. O objetivo principal da pesquisa descrita consiste na identificação de problemas no processo inerente à criação e manutenção de um tesauro, tanto decorrente da falta de conhecimento da terminologia por alguns lexicógrafos, como também da dificuldade em se criar um tesauro devido a fenômenos lingüísticos. Como fonte de teste, utilizou-se o projeto *MorphoSaurus*, que possuía 86 versões do tesauros arquivadas na forma de bases de dados já modificadas, cobrindo um período de nove meses. A partir dessas bases de dados, criou-se um programa para registrar as alterações realizadas e identificar as anomalias detectadas. Os resultados foram comparados com aqueles obtidos por meio de outra abordagem baseada em *corpora* comparáveis. A anomalia de relacionamento ocorreu 146 vezes, tendo havido 76 repetições distintas e dessas, 27 foram encontradas na abordagem de *corpora* comparáveis. Quanto à anomalia de delimitação, não se encontrou nenhuma nesse período. Com relação à anomalia de permanência, foram encontradas 5, as quais também apareceram na abordagem de *corpora* comparáveis. Finalmente, a anomalia de tipo apresentou 18 ocorrências, sendo que todas também foram reconhecidas na abordagem de *corpora*. Neste caso, constatou-se que o registro de procedimentos auxilia efitivamente no refinamento do léxico, pois o sistema *MorphoSaurus* já possui quantidade significativa de classes representativas.

Palavras-chave: Recuperação de Informação, Tesauro, Lexicografia, Indexação de Informação, Controle de Qualidade.

ABSTRACT

To build a thesaurus with synonyms-based classes and include acceptions in medical domain is not a trivial task, due to the inherent terminology complexity. In this dissertation we show some highlighted problems present in the manual mode thesaurus construction and describe an approach to help the lexicographers in its construction and maintenance process. We introduce the conception of thesaurus management anomalies as sequence of such actions done by the thesaurus curators that consume effort without any positive impact on the quality of the thesaurus. In this research 4 kinds of frequently incorrect procedures were mapped on the thesaurus: relationship, type, delimitation and permanence anomalies. The main goal of this research is the identification of inherit problems in the process of thesaurus' construction and maintenance due to the lack of lexicographers' knowledge as well as the difficulty to solve linguistics phenomena. As workbench we used the last versions of *MorphoSaurus* Project's thesaurus representing 86 modified data bases covering a period of nine months. In order to register the procedures and to identify the anomalies of the 4 kinds of incorrect procedures, a program as implemented. The obtained results were compared to those performed with another approach based on comparable corpora. The relationship anomaly did occur 146 times from which 76 different repetitions and of those, 27 were found in the corpus-based error approach. We did not find any occurrence of delimitation anomaly during the considered period. Also, related to the permanence anomaly we had found 5 occurrences which were coincident to the corpus-based error approach. Then, type anomaly had presented 18 occurrences, and all of them were recognized by the corpus-based approach. The thesaurus of *MorphoSaurus* Project has a consolidated coverage. So, after the analysis of the anomalies we concluded that the verification of procedures register contributes effectively to the thesaurus refinement.

Keywords: Information Retrieval, Thesaurus, Lexicografy, Information Indexing, Quality Control.

CAPÍTULO 1

INTRODUÇÃO

1.1. CARACTERIZAÇÃO DO PROBLEMA

Há muito tempo, a humanidade produz, armazena e organiza as informações para serem recuperadas quando houver necessidade (CARVALHO, 1999). Devido à intensa dinamicidade e à enorme quantidade de conhecimentos gerados em múltiplas áreas, emerge paralelamente ao crescimento exponencial de informações, a necessidade de gerenciá-las de forma mais eficaz e de disponibilizar de forma mais eficiente documentos relevantes, quando assim o usuário necessitar. É humanamente impossível processar toda a gama de informações disponíveis, em sua grande maioria de forma digital (SARACEVIC, 1996).

Diante desse crescimento desordenado de informações, em alguns casos, de forma desorganizada e banalizada, surgiu a necessidade tanto de aperfeiçoar técnicas existentes como também de criar outras mais sofisticadas, no sentido de melhorar a recuperação de documentos com ênfase na busca dos que satisfaçam com maior precisão os requisitos do usuário, ou seja, a busca de documentos relevantes (RIJSBERGEN, 1979; SPARCK-JONES e WILLET, 1997; KOWALSKI, 1997; CROFT, 2000; MEADOW, 2000).

A Internet é vista como uma grande fonte de informação e cresce de forma incessante (CHANKRAVARTHY e HAASE, 1995). Para procurar informações relevantes dentro dessa fonte, é necessário o emprego de técnicas, as quais são desenvolvidas na área de Sistemas de Recuperação de Informação (SRI) (YATES, 1996).

As ferramentas de busca tornam-se cada vez mais indispensáveis tanto na Internet como em ambientes *Intranet*, utilizando cada vez mais técnicas avançadas para recuperação de informação (HERSH, 1996). Apesar da grande disponibilidade de ferramentas de busca, alguns problemas podem ocorrer e

dentre eles, o emprego inadequado de ferramentas de busca por grande parte dos usuários, seja referente ao assunto abordado, à tecnologia empregada na ferramenta ou a sua utilização (WILLIE e BRUZA, 1995).

A área de Recuperação de Informação (RI) textual pode ser classificada como RI monolíngüe ou RI multilíngüe (*cross-language*) (OARD, 1997). A diferença entre RI multilíngüe e monolíngüe é a habilidade do sistema multilíngüe recuperar documentos em uma língua natural diferente da utilizada na consulta.

Existem basicamente dois processos envolvidos na Recuperação de Informações: a indexação e a 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, utilizado para representar conteúdos de documentos, com a finalidade de classificação ou busca de informação (CINTRA et al., 2002).

A idéia principal de se utilizar um tesauro é prover um vocabulário controlado de referência a um SRI (FOSKETT, 1997), com o objetivo de melhorar a qualidade de recuperação. Por intermédio desses tesauros, pode-se indexar e recuperar documentos em um determinado domínio.

A construção de um tesauro envolve alguns passos, mas normalmente começa-se definindo o domínio de atuação. Uma vez definidos e delimitados os limites desse domínio, o passo seguinte é compilar um *corpus* de termos da terminologia que seja representativa desse domínio, de tal forma que sirva de matéria prima para a construção do tesauro proposto (SOERGEL, 1997 a). A importância de usar um tesauro decorre do fato de que grande parte da informação é criada e expressa por meio da linguagem natural. Isto acontece porque a linguagem natural representa o modo de comunicação dos seres humanos, onde se utilizam vocabulários diferentes para expressar as suas intenções (FURNAS, 1987).

Todo ramo do conhecimento humano necessita instituir sua própria terminologia, necessária na sua forma de se comunicar e expressar, existindo uma variedade inevitável de terminologias no meio profissional (KRIEGER, 2006). A medicina, como uma das mais antigas ciências, desenvolveu sua própria linguagem que, ao leigo, mostra-se de difícil entendimento. Da mesma forma, o estudante de medicina, no início do curso, espanta-se com tantos

termos novos que deve aprender e cujo significado é de difícil assimilação.

Na área de saúde, a terminologia empregada é caracterizada por formas complexas de composição, derivação e inflexão, assim como pela geração contínua de novos acrônimos, abreviações e nomes próprios. Além disso, existe o fato de que nem sempre documentos relevantes estão na língua nativa do usuário (SCHULZ e HAHN, 2000).

Construir um tesauro na área médica com a idéia de reunir classes de sinônimos e contemplar acepções não é uma tarefa trivial devido à complexidade inerente à própria terminologia, dentre as quais pode-se citar (SCHULZ et al., 2000 a):

(1) variação ortográfica: *diabetes mellitus, diabete melito*;

(2) derivação: *diabetes, diabéticos, diabéticas, antidiabéticas*;

(3) composição: *hiperprebetalipoproteinemia*;

(4) sinônimos: *nepho, renal / estômago, gastr*;

(5) acrônimos: *AVC, ECG, DPOC, SIDA*,...;

(6) nomes próprios: *Diclofenaco, Viagra, Parkinson*,...

Somam-se ainda, aspectos de ordem semântica e conceitual, relevância lexical e semântica, etc... A decisão a ser tomada com relação à equivalência de dois conceitos como sinônimos ou relacionamento a uma possível acepção, na maioria das vezes, também não é uma tarefa fácil. Em alguns casos, depende de consenso entre os lexicógrafos devido aos aspectos inerentes ao próprio processamento da linguagem natural, tais como: comunicação humana, sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos (FRANCONI, 2001).

Criar um tesauro manualmente é um trabalho bastante oneroso e difícil de ser realizado (NIE et al., 1999), pois demanda tempo e recursos humanos especializados. Outro problema é a individualidade, pois o entendimento do significado dos termos depende muito de visões e de contextos individuais.

Portanto, é necessário rigor no exame dos elementos selecionados com o objetivo de produzir um adequado desempenho num sistema de recuperação de informação, assim como o monitoramento da própria construção e sua manutenção. Dessa forma, para assegurar a representatividade e efetividade do tesauro de um domínio específico, é necessário controle tanto do vocabulário, o qual possa expressar as idéias, quanto monitoramento das

ações dos lexicógrafos sobre o tesauro.

Esses problemas geram uma demanda enorme na construção e manutenção do tesauro, haja vista que normalmente emprega-se uma equipe multidisciplinar com interpretações, por vezes diferentes, sobre como tratar determinada classe ou conjunto de classes de conceitos. Diante desse fato, surge a necessidade de mapear os procedimentos realizados, de forma a gerar parâmetros que possam servir de apoio no direcionamento dos procedimentos a serem realizados.

Desse modo, mapeou-se 4 tipos de procedimentos freqüentes no tesauro: anomalias de relacionamento, tipo, delimitação e permanência.

Os lexemas constituem as classes de equivalência (CEq). As CEq denotam um sentido em comum, isto é, correspondem aproximadamente ao que se entende por conceito.

Na delimitação semântica, a definição desempenha papel importante na formação e organização dos termos, afetando a estruturação do campo conceitual da linguagem. Nem sempre, porém, pode-se contar com definições claras, seja pela falta de dicionários técnicos que auxiliem o trabalho de organização das linguagens, seja pela dificuldade de delimitação das classes de conceitos ou delimitação semântica. Esse último aspecto explica, por exemplo, o fato de que a tarefa de organização de termos nas áreas das humanidades seja mais complexa. Em face desses problemas, é vital conhecer os fatores envolvidos no processo da criação da definição, bem como suas conseqüências para o tratamento da informação (LARA, 2004).

Ainda com relação à delimitação semântica, é possível que não exista consenso sobre a forma como deva ser feita. Essa formulação pode parecer provocadora, porém, ela traduz nada menos que as dificuldades reais que os pesquisadores encontram (VERSTRAETE, 2001).

Contudo, a dimensão semântica fornece a chave decisiva para identificar a unidade léxical no discurso. A semântica é o que define e traz as informações para que se perceba o segredo inerente ao significado de cada palavra (BIDERMAN,1999), pois ela está relacionada ao significado, não só de cada palavra, mas também do conjunto delas resultante (SAINT-DIZIER, 1999).

A falta de delimitação precisa de um termo dentro de um conceito, ou delimitação conceitual do termo, representa a relação de equivalência, o que

significa dizer que os termos dentro de uma classe são sinônimos lexicais representando conceitos iguais (STREHL, 1998).

A delimitação do termo é obtida a partir da definição, mas não a definição lexicográfica, tal como ocorre num dicionário de língua, mas a definição dentro de um campo conceitual (STREHL, 1998).

A referência dos termos está no objeto, não como realidade física, mas como realidade cultural, construída em função de determinados objetivos. A definição estabelece um sentido (DAHLBERG,1978).

A falta de delimitação precisa dos termos dentro de uma classe de equivalência cria um problema que influencia na recuperação de documentos, podendo acontecer dois casos**:** a equivalência de termos que não são sinônimos (o que resulta em queda do parâmetro *precision*), ou a falta de equivalência entre termos que são sinônimos (o que resulta em queda do parâmetro *recall*).

O conceito de relevância é estudado por cientistas da informação e da semântica. A ciência da informação tem tentado entender e explicar o fenômeno da relevância através de diversas abordagens, mas ainda existem várias falhas, como as que se relacionam com valores humanos envolvidos na comunicação do conhecimento (FIGUEIREDO, 1977).

Um lexema irrelevante equivale a um lexema pouco freqüente, que não é informativo ou não tem influência suficiente para contribuir com a informação útil devido a sua baixa ocorrência (APTÉ et al., 1994).

Um outro aspecto a ser considerado relacionado à representação de um conceito por descritor semântico, aliado à de identificação de termos semanticamente relevantes, trata-se da delimitação do tamanho de uma *string*. Por exemplo, a delimitação do tamanho de um termo simples (não composto), de um lexema, no projeto *MorphoSaurus*, é guiado por motivos funcionais além dos gramaticais, com o objetivo de gerar descritores semânticos corretos após a segmentação de uma palavra. Se for necessário acrescentar ou retirar uma letra da seqüência de caracteres para manter a integridade da representação de um conceito na linguagem artificial, isso é feito.

O trabalho desenvolvido busca analisar os 4 tipos de problemas descritos, tornando-se uma ferramenta de auxílio aos profissionais de saúde para inclusão de morfemas na criação e manutenção de um tesauro médico.

1.2. OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral dessa dissertação é a criação de uma metodologia com o intuito de monitorar o processo inerente à criação e manutenção de um tesauro, verificar a existência de problemas decorrentes tanto da falta do total domínio da terminologia pelos lexicógrafos como também da dificuldade em criar um tesauro devido a fenômenos lingüísticos. Por intermédio da análise dos procedimentos tomados pelos lexicógrafos, pode-se identificar anomalias resultantes dessas modificações no tesauro.

1.2.2 Objetivos Específicos

Especificamente, inserem-se como objetivos específicos:

(1) verificar se existem anomalias de relacionamento (An rel), analisando os procedimentos nas relações dentro do tesauro, executados pelos lexicógrafos, e compreendendo as mudanças realizadas envolvendo a criação, mudança ou eliminação de relacionamentos entre classes (*has_sense* e *has_word_part*);

(2) verificar ocorrências de anomalias de tipo (An tip), baseadas no processo de equivalência, analisando as equivalências realizadas pelos lexicógrafos;

(3) verificar se ocorrem anomalias de delimitação (An del) de um lexema, analisando o processo de edição do léxico executado pelos lexicógrafos;

(4) verificar a existência anomalias de permanência (An per), analisando os procedimentos de inclusão e exclusão de lexemas no tesauro realizados pelos lexicógrafos.

1.3. ESTRUTURA DA DISSERTAÇÃO

No capítulo 2, apresenta-se a fundamentação teórica para o embasamento deste trabalho, e são esclarecidos assuntos referente à Formação da Linguagem com os seus sub-capítulos linguagem natural, Normalização Lingüística e Ambigüidade. Em seguida, abordam-se os temas Tesauro, RI, Composição de um SRI e Técnicas de Avaliação de Performance. Logo depois, caracteriza-se o Sistema *MorphoSaurus*, o qual foi utilizado como fonte de teste para essa dissertação.

O capítulo 3 contém a metodologia aplicada nesse trabalho, para identificar os processos anômalos que ocorreram, apresentando a metodologia para coleta de dados, identificação das anomalias e validação, no momento de sua criação e manutenção pelos lexicográficos.

O capítulo 4 apresenta os resultados obtidos, categorizando os problemas em quatro classes de anomalias e comparando com a lista de discussão dos lexicógrafos.

No capítulo 5, apresenta-se a discussão da metodologia criada e dos resultados obtidos e, por fim, no capítulo 6, as conclusões, com as considerações finais sobre a pesquisa.

CAPÍTULO 2

ESTADO DA ARTE

2.1. FORMAÇÃO DA LINGUAGEM

2.1.1 Linguagem Natural

A linguagem natural (LN) representa a maneira mais simples e natural de entendimento dos seres humanos. Assim, sendo a interação entre interlocutores o princípio que fundamenta a linguagem, é possível afirmar que na LN podem ser usadas expressões-padrão para lembrar uma entidade que está presente no contexto, ou é de conhecimento dos envolvidos. Ou seja, duas relações entram em jogo: a relação direta entre a própria pessoa e a dela com a sociedade. Os discursos produzidos, determinados por coerções sociais, exigem interpretação, contando com julgamentos de valor, experiências vividas e trocadas, entre outras questões. O ato interpretativo é, então, fundamental para proporcionar o relacionamento de figuras lingüísticas com objetos do mundo, ou seja, é o entendimento que se faz crucial para o desenvolvimento de aplicações de LN mais confiáveis (OLIVEIRA, 1999).

Dessa maneira, normalmente os documentos para realização de pesquisa são representados em LN, pois é o meio mais comum de comunicação humana. No entanto, essa ação via LN apresenta problemas. Em geral, os sistemas que implementam este estilo trabalham com um subconjunto da linguagem, suportá-la em toda sua abrangência é complicado, pois há milhares de termos existentes e novos termos são criados, além das alterações dos termos antigos. A LN tem uma capacidade de expressão enorme e é exposta de maneira detalhada, individualizada e espontânea, podendo ser: ambígua, imprecisa, redundante e imprevisível.

Sendo assim, a utilização da LN nesses sistemas deve ser feita com

cautela, pois no ato comunicativo as pessoas podem utilizar vocabulários diferentes para expressar a mesma intenção e, como conseqüência, podem usar termos diferentes para referenciar-se ao mesmo conceito ou significado. Técnicas complementares devem ser empregadas para minimizar tais problemas (FURNAS, 1987).

Para reduzir tais fatores agravantes, Lewis (1996), sugeriu o uso de LN restrita, de forma que a representação da LN seja realizada em formatos padrões que fossem permitidos a sua utilização somente dentro desta linguagem.

Por aspirar à construção de uma linguagem perfeita, que não apresenta as variações e peculiaridades da LN, Frege (1982) deixou de fora de sua teoria aquilo que não pode ter uma referência no mundo, pois parte do princípio de que a LN apresenta ambigüidades que poderiam ser eliminadas numa linguagem artificial e objetiva.

A idéia principal do ponto de vista computacional e do usuário no momento da busca da informação seria uma consulta de forma exata, compacta, consistente e processável, o que a LN, muitas vezes, não possui.

2.1.2 Normalização Lingüística

O reconhecimento de variações lingüísticas encontradas internamente em um texto possibilita o controle do vocabulário (JACQUEMINe TZOUKERMANN, 1997). A normalização lingüística pode ser dividida em três âmbitos: morfológico, sintático e léxico-semântico (ARAMPATZIS et al., 2000).

A normalização morfológica ocorre quando há redução dos itens lexicais através de sua fusão, que procura representar classes de conceitos. Essa fusão é a operação que combina a representação de dois ou mais termos em um único conceito, reduzindo a uma única forma conceitual.

Os procedimentos mais conhecidos para geração de termos candidatos à equivalência são:

(1) *stemming*: reduz uma palavra à sua parte significante semelhante ao seu radical (*stem*) através da eliminação de afixos oriundos de derivação ou de

flexão (ORENGO e HUYCK, 2001). Por exemplo: *stemming* (pulmão) = *stemming* (pulmões) = *pulm*;

(2) *lematização* (“*lemmatization*”): reduz os adjetivos à forma masculina singular, os substantivos à forma singular, e os verbos na sua forma infinitiva (ARAMPATZIS et al., 2000). Por exemplo: lematização (encaminho) = lematização (encaminhar) = encaminhar.

A normalização sintática ocorre quando há a normalização de frases semanticamente equivalentes em uma forma única e representativa das mesmas, como “cirurgia de fígado e pulmão” e “cirurgia de pulmão e fígado”.

A normalização léxico–semântica ocorre quando são empregados relacionamentos semânticos, e pode-se considerar como relacionamento à sinonímia entre os itens lexicais, para criar um agrupamento de similaridades semânticas, identificado por um item lexical que representa um conceito único (JACQUEMIN e TZOUKERMANN, 1999).

Pode-se encontrar duas formas de normalização lexical. De um lado está a normalização morfológica através do processo de *stemming*, que explora similaridades morfológicas. No outro extremo, encontra-se a normalização léxico-semântica, por exemplo, através de busca de sinônimos em um tesauro, considerando informações terminológicas (JACQUEMIN e TZOUKERMANN, 1999).

2.1.3 Ambigüidade

A ambigüidade é a propriedade que faz com que um objeto lingüístico, seja ele uma palavra, um termo composto ou um texto completo, possa ser interpretado de formas diferentes (FUCHS, 1987). Quanto ao nível de processamento, existem dois tipos de ambigüidade: sintática e semântica (JURAFSKY e MARTIN, 2000).

A ambigüidade sintática ocorre quando um item lexical pode pertencer a mais de uma classe gramatical, como “almoço” que pode ser substantivo ou verbo (SMEATON, 1997).

A ambigüidade semântica ocorre quando uma única palavra possui mais

de um sentido, toma-se como exemplo a diversidade de significados da palavra passar dentro de um dado contexto, como: “passar a ferro”, “passar no exame” e “passar em casa” (BEARDON e HOLMES, 1991).

As causas da ambigüidade podem ser, segundo Beardon e Holmes (1991), dos seguintes tipos: lexical, quando uma palavra pode ter vários significados ou estrutural, neste último caso existe a possibilidade de existir mais de uma estrutura sintática para uma sentença, podendo ser: local, quando a ambigüidade pode ser resolvida não levando em consideração o contexto empregado e ainda, pode ser global, quando é imprescindível a análise contextual para solução.

Um exemplo de ambigüidade local seria: “ele cuidou do computador com esperança”, o sentido “computador com esperança” em princípio pode ser descartada. E, em: “ele cuidou do paciente com esperança”, há ambigüidade estrutural global, pois é possível construir duas associações diferentes: “cuidou com esperança” e “paciente com esperança” (BEARDON e HOLMES, 1991).

Na ambigüidade lexical podem ocorrer dois fenômenos lingüísticos: a homonímia e a polissemia (KROVETZ, 1997; KROVETZ e CROFT, 1992).

A homonímia consiste na relação entre duas ou mais palavras que, apesar de possuírem significados diferentes, possuem a mesma estrutura fonológica. Esses termos podem ser classificados em homógrafos, referente à escrita como, por exemplo: almoço, seco (substantivo) e almoço, seco (verbo); e podem ser homófonos, quando a diferença é referente ao som como em cassar e caçar, ou ainda, podem ser homônimos perfeitos, quando o termo é um homófono e homógrafo ao mesmo tempo, como por exemplo, verão, e são (SANTOS, 2002).

Já na polissemia, a palavra pode adquirir diferentes significados: toma-se, por exemplo, a palavra alta, que pode significar altura, tratando-se de uma pessoa alta, ou de alta hospitalar ou ainda, de pressão alta, entre outros tantos significados.

No contexto da RI, Krovetz (1997) defende três hipóteses relacionadas à ambigüidade lexical:

(1) a resolução da ambigüidade lexical beneficia o desempenho da recuperação de informação;

(2) os significados das palavras determinam uma separação entre os

documentos relevantes e não relevantes;
(3) mesmo em um *corpus* pequeno e de domínio específico, há uma proporção significante de ambigüidade lexical.

As informações provenientes de dicionários como morfologia, categoria gramatical e composição de termos são fontes de evidência para a resolução de ambigüidades (KROVETZ, 1997).

2.2. TESAURO

Tesauro é um modelo léxico-semântico de realidades conceituais, expressas na forma de um sistema de termos e suas relações, que oferece acesso por diferentes vias e é usado como ferramenta de processamento e busca de uma unidade de RI (MILLER, 1997).

A palavra “tesauro” tem origem etimológica do latim *thesaurus*, que se originou do grego *thesaurós*, que vem sendo utilizada para designar um “tesouro de palavras” ou “armazém de palavras” ou ainda “repositório de palavras”, pois este tipo de dicionário deve fornecer “riqueza” em conceitos e respectivas relações semânticas, de forma a ter uma grande abrangência em um determinado domínio de conhecimento. Tesauro também significa vocabulário, dicionário ou léxico.

Então, tesauro é um sistema de vocabulário baseado em conceitos, com inclusão de termos preferidos chamados de descritores, termos não preferidos ou não descritores e suas inter-relações. Aplica-se a um ramo do conhecimento, onde este vocabulário que tem por função controlar a terminologia utilizada para indexar e recuperar documentos (MOTTA, 1987).

Um tesauro pode contemplar um domínio de conhecimento específico, que seria um tesauro especialista, desenvolvido com termos de determinada especialidade ou mesmo ser genérico, que abranja todos os assuntos. Gonzales (2001) referiu que os tesauros genéricos são normalmente criados manualmente, enquanto que a criação automática destes envolve o desenvolvimento de modelos de tesauros sobre um domínio específico.

Tesauros têm sido utilizados para indexar e recuperar informação em

diversos domínios. Além disso, têm como função fornecer um vocabulário constante para indexação da informação, e permitir aos usuários utilizarem-no de forma intuitiva e organizada, além de pesquisar informações de seu interesse, e ainda permite que a consulta em alguns casos, seja em mais de uma língua, por exemplo, no *General European Multilingual Environment Thesaurus* (GEMET, 2005).

Além de seu uso para indexação de assuntos, o tesauro pode ainda oferecer outros recursos, através da exploração das relações entre seus termos, através de notas de escopo, ou outras informações, tal como a origem do termo.

Nesse contexto, a eficiência do uso de tesauros tem sido comprovada em estudos que demonstram ganhos de precisão nas consultas da ordem de 30% (SILVEIRA, 2003) e seu uso tem se difundido, passando da indexação de acervos de bibliotecas por meio de fichas catalográficas em papel, até a indexação de acervos multimídia digitais. É importante ressaltar que a forma de elaboração do tesauro vai influir na sua eficiência. Isto vai de encontro aos métodos empregados na construção e manutenção de cada tesauro.

Vocabulário controlado, utilizado em SRI, procura minimizar essas características da LN, utilizando tesauro para restringir o vocabulário de indexação e de consulta, de forma que uma idéia possa ser expressa somente de uma única maneira.

A utilização de vocabulário controlado está ligada à utilização de tesauro. Estas técnicas buscam indexar documentos com o uso de índices que representem conceitos únicos.

O tesauro tem por função apoiar na área de documentação, organização do vocabulário de indexação e recuperação, podendo ser utilizado em ambientes organizacionais, na representação de assunto dos documentos, na busca de informação e também no processo de classificação pelo indexador (CURRÁS, 1995).

O tesauro tem uma função importante em um SRI, por determinar:

(1) os termos que podem ser usados em um SRI;

(2) os termos que podem ser usados em uma busca, para que esta retorne documentos relevantes e

(3) uma estrutura que permite a introdução de novos termos e relações de

modo a aproximar a linguagem do usuário ao do sistema e realizar alterações dos termos existentes.

O tesauro pode ser dividido em 3 tipos: em função da língua, do nível de especificidade e do assunto que cobrem (GOMES, 1990).

Quanto à língua, podem ser monolíngües ou multilíngües, envolvendo dois ou mais idiomas.

Quanto à especificidade, dividem-se em:

(1) microtesauros, onde os descritores denotam conceitos em um nível maior de especificidade e se referem a um domínio mais restrito e

(2) macrotesauros, onde os termos representam conceitos mais ou menos amplos, número menor de descritores. É composto por vários microtesauros, relacionados entre si por referência cruzada, abrangem um maior número de assuntos, sendo cada assunto um microtesauro especialista.

Quanto ao escopo ou assunto, há (1) os projetados para atuar em um problema: um tesauro multidisciplinar, pois envolve descritores das várias áreas relacionadas com o problema; e aqueles (2) voltados para um único assunto, que é o caso de tesauro voltado para uma única disciplina como medicina.

2.3. RECUPERAÇÃO DE INFORMAÇÃO

2.3.1 Visão Geral

A Recuperação de Informação (RI) é um campo vasto e complexo, que engloba técnicas para construções de sistemas que vão de formas de como representar informações até como acessá-las. Essas técnicas não devem se concentrar nos dados por eles manipulados, mas sim nas informações contidas nesses mesmos dados. O grande desafio desses sistemas está em responder, da melhor forma possível, consultas de informações feitas por seus usuários (SMEATON, 1997).

A recuperação, a representação, o armazenamento, a organização e o

acesso são processos de gestão da manipulação da informação (YATES e RIBEIRO, 1999).

A RI tem como função recuperar documentos dos mais variados tipos, em meio a uma grande quantidade de documentos desorganizados, partindo do princípio de uma necessidade de informação (THE FREE DICTIONARY, 2005).

As bibliotecas foram os primeiros usuários deste sistema e, no começo, os SRI não passavam de uma evolução dos catálogos de livros. Com o surgimento da Internet e das bibliotecas digitais, este campo acabou se tornando um desafio, pois na Internet as informações são dinâmicas e totalmente descentralizadas.

O procedimento em recuperar informações pode ser uma experiência frustrante para alguns usuários. O universo da informação é dinâmica, pois os documentos são alterados freqüentemente. Além disso, cada usuário possui necessidades diferentes, como também conhecimentos diferentes, tanto com relação à informação quanto à utilização da web como fonte de pesquisa (WANG, 2000).

Um SRI é formado por uma coleção de documentos previamente catalogados, um mecanismo de busca e uma interface com o usuário. A partir de uma consulta realizada pelo usuário, o mecanismo de busca é acionado que, então, realiza uma busca de documentos dentro da coleção que satisfaçam a consulta realizada.

Esta é a forma básica de um modelo de RI. Cada SRI possui sua peculiaridade, de forma que existem vários processos diferentes para cada SRI, processos estes que existem para buscar a melhor forma de atender a necessidade do usuário.

O objetivo d SRI num cenário é compartilhar os documentos de maneira rápida e fácil, deixando fluir o conhecimento do negócio. Um SRI eficiente torna-se uma vantagem competitiva para as corporações, na medida em que aumenta a sua produtividade (MILSTEAD, 1998).

2.3.2 Composição de um Sistema de Recuperação de Informação

Um SRI foca a extração de informações, que nada mais é do que recuperar documentos textuais com qualidade, selecionando e priorizando estes documentos de forma a serem relevantes ao usuário (KUSHMERICK e THOMAS, 2003). Para isso, trabalha-se com a organização, o armazenamento e o acesso aos itens que compõem uma biblioteca digital.

A figura 1 apresenta um modelo básico de RI. A constituição da coleção de documentos é o primeiro passo na criação de um SRI, que pode ser composta por um conjunto de documentos de interesse de uma comunidade especifica de usuários ou de documentos das mais variadas fontes de texto, como aquelas formadas a partir da Internet. Tendo uma coleção de textos, é necessário que um SRI seja capaz de organizado-los e armazenado de forma a facilitar o acesso a esses documentos. Para que isso aconteça, há a necessidade de um índice.

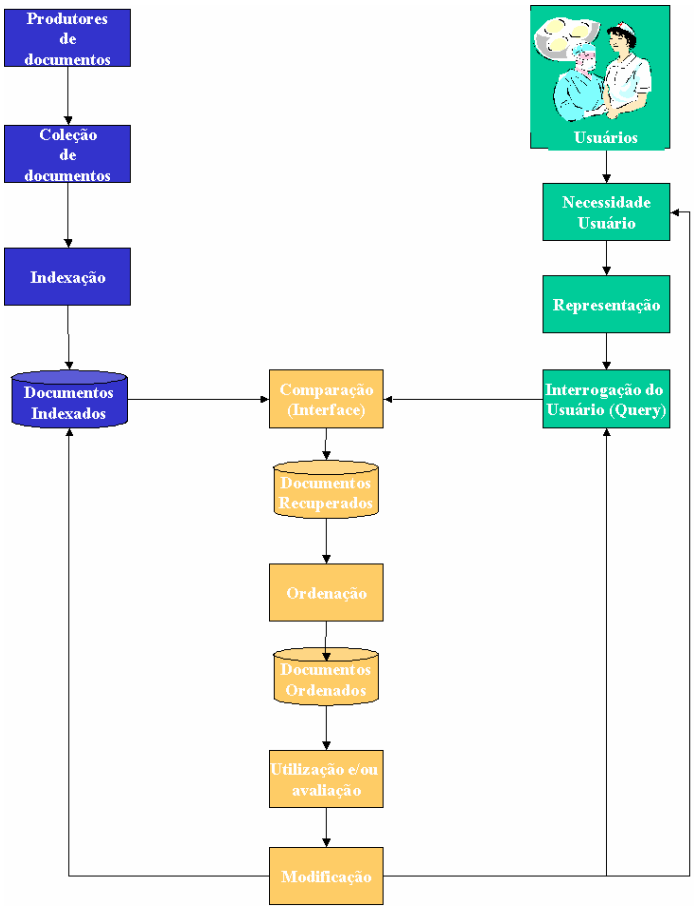


Figura 1 – Forma básica de um modelo de Recuperação de Informação.

Nos processos manuais, os índices são analisados através do conteúdo de cada documento**.** O avaliador seleciona os termos que representam o documento para utilizar como índice. Os vocabulários controlados são ferramentas utilizadas na formação dos índices, pois são utilizadas na tradução do conteúdo dos documentos, com o objetivo de padronização da linguagem de indexação.

O projetista de um SRI deve decidir se haverá controle ou não de vocabulário na indexação na pesquisa. Existem quatro maneiras de se utilizar vocabulários controlados (LANCASTER, 1993):

(1) controlar o vocabulário na indexação (entrada) e na pesquisa (saída). Exemplo *MEDLINE* com o *MeSH*;

(2) controlar o vocabulário na indexação e não controlar na pesquisa. Exemplo *MEDLINE* com o *PUBMED search*;

(3) não controlar a indexação, mas sim a pesquisa, ou seja, utilizar tesauro somente na pesquisa;

(4) não exercer qualquer controle, o que caracteriza um SRI de LN, como os sistema que alimentam os motores de busca na Web.

Nos SRI computacionais, o índice é gerado por algoritmos que analisam sintaticamente os documentos e extraem os termos de indexação para compor o índice. Este processo consiste em associar termos de indexação a documentos, assim, melhorando o retorno de documentos que satisfarão as necessidades do usuário, aumentando a eficiência da recuperação da informação (FUHR e BUCKLEY, 1990).

Antes da indexação, é necessário realizar um pré-processamento para não criar índices com palavras ou termos que não tenham função dentro do texto. Nesse pré-processamento acontecem operações tais como para redução das palavras a sua raiz gramatical (*stemming*), exclusão de acentos, hífens, espaços em branco e exclusão de palavras sem valor de indexação (*stopwords*) (YATES e RIBEIRO, 1999). Em tais operações de transformação, a estrutura da língua é perdida e também parte da semântica contida nos documentos.

Após esses processos, realiza-se a comparação entre a interrogação do usuário e os documentos indexados. Como resultado desse processo, retira-se o conjunto de documentos que freqüentemente são ordenados, conforme a sua

relevância. Cada SRI define a sua técnica de ordenação que classifica a disposição do retorno desses documentos.

A relevância é um aspecto subjetivo e é definido pelo usuário, variando conforme o tempo de utilização e o usuário, sendo complexo definir qual a melhor função de relevância. Um dos aspectos mais importantes dos SRI é definir uma boa função de relevância (BELKIN e CROFT, 1992).

Os vocabulários controlados permitem recuperar parte da semântica perdida, empregando classes de conceitos e seus relacionamentos. Essas operações são executadas para todos os documentos, a fim de obter as palavras-chave que formam o vocabulário da coleção. Para cada palavra-chave, o SRI gera o conjunto de documentos, onde a palavra-chave ocorre na forma de lista invertida, também chamada de arquivo invertido (YATES e RIBEIRO, 1999).

O arquivo invertido é uma lista de termos ou palavras ordenadas, onde cada um desses termos possui *links* para os documentos (YATES, 1992), conforme representado na figura 2.

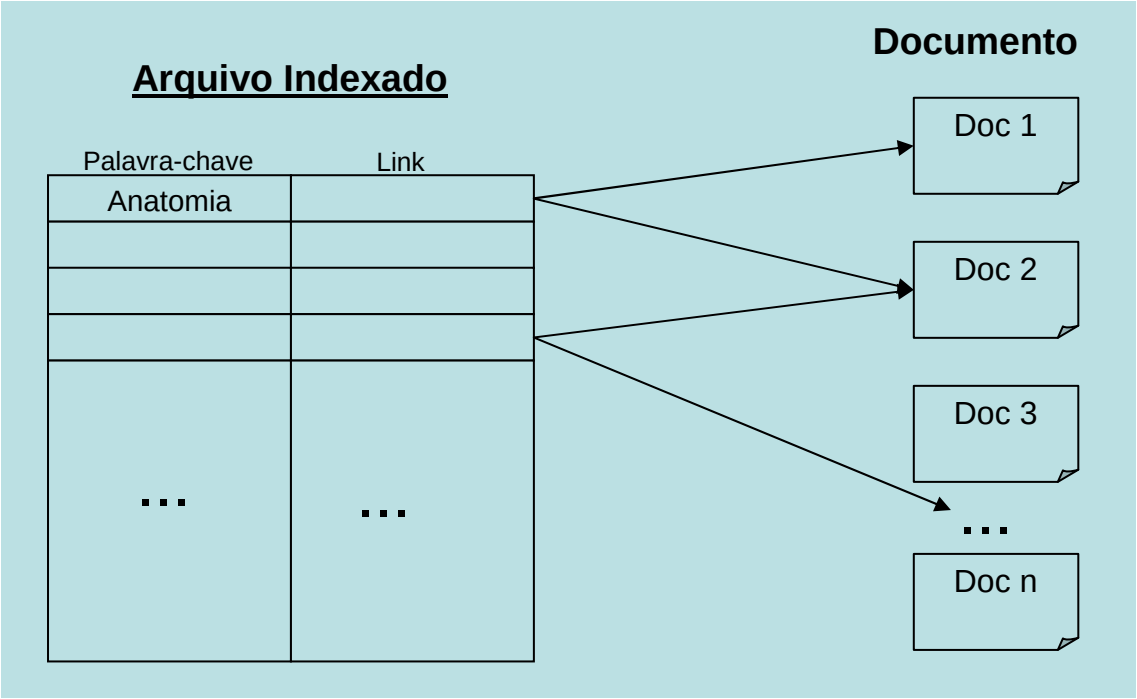


Figura 2 – Arquivo invertido utilizando array ordenado (YATES, 1992).

Devido à sua rapidez de acesso e à sua facilidade de identificação de documentos relevantes a um termo, essa estrutura é uma das mais utilizadas

em SRI (KOWALSKI, 1997).

Nos SRI, o sistema de indexação pode suportar tesauros manuais ou automáticos. Por meio desses tesauros, ocorre a indexação dos documentos que, por sua vez, a partir de suas equivalências e relações, podem expandir ou restringir a pesquisa dos usuários.

2.3.3 Técnicas de Avaliação de Performance

As duas principais medidas de avaliação de performance utilizadas em SRI são precisão (*precision)* e revocação (*recall)*, e foram primeiramente propostas por Kent et al. (1955).

Na precisão, avalia-se a quantidade de documentos relevantes dentro de um conjunto retornado por uma determinada consulta, ou seja, mede-se a capacidade do sistema de recuperar somente documentos relevantes (BAEZA-YATES e RIBEIRO-NETO, 1999; VOORHEES e HARMAN, 1996; HARMAN, 1991). Assim, é calculado pela equação 1:

$$P = Rc / T \qquad \text{(equação 1)}$$

Onde P representa precisão, Rc é o total de documentos relevantes recuperado e T o total de documentos retornados em uma determinada consulta. No exemplo da figura 3, tem-se a precisão de 50%.

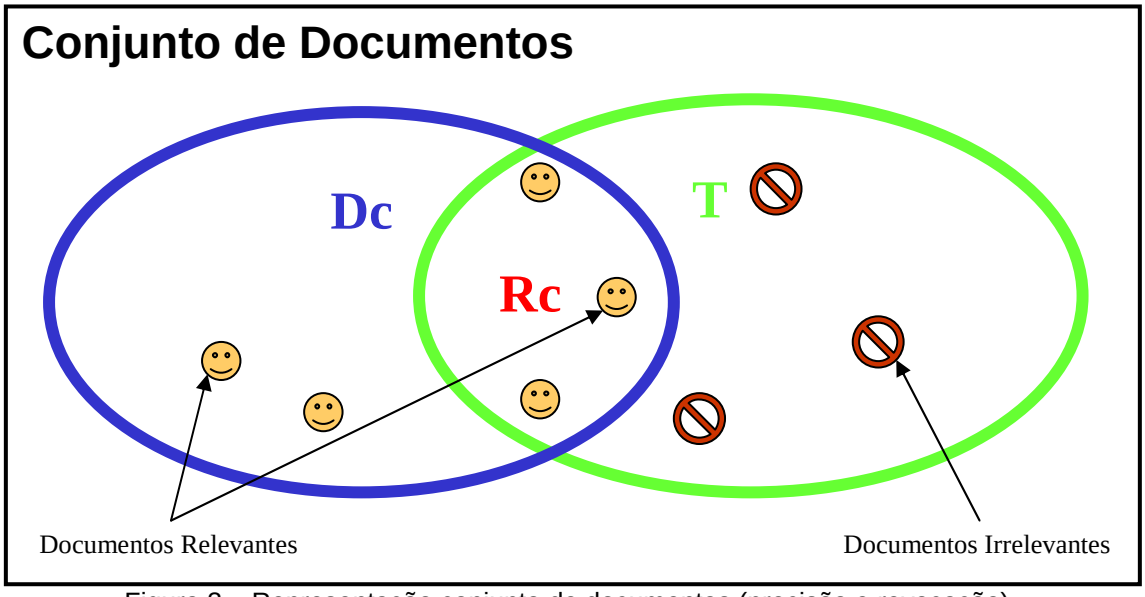


Figura 3 – Representação conjunto de documentos (precisão e revocação).

Na revocação, avalia-se a porção dos documentos relevantes de toda a coleção que foram retornados na consulta, ou seja, é uma medida de capacidade do sistema de recuperar todos os documentos relevantes (BAEZA-YATES e RIBEIRO-NETO, 1999; VOORHEES e HARMAN, 1996; HARMAN, 1991). É calculado pela equação 2:

R = Rc/Dc (equação 2)

Onde R corresponde à revocação e Dc representa o total de documentos relevantes dentro do conjunto. Conforme na figura 3, o resultado desta equação de Revocação é igual a 60%.

A idéia de realizar a avaliação com base nesta medida visa focar a avaliação nos documentos efetivamente observados pelo utilizador (SILVERSTEIN, 1999).

Para obter uma precisão alta, o número de documentos recuperados normalmente é menor, fazendo assim com que o número de documentos relevantes seja baixo e, desse modo, diminuindo a revocação. Normalmente, busca-se um equilíbrio em a precisão e a revocação.

2.4. CARACTERIZAÇÃO DO SISTEMA *MORPHOSAURUS*

2.4.1 Visão Geral

Tendo em vista que muitos documentos relevantes encontram-se em diferentes línguas e não somente na língua nativa do usuário, há uma deterioração da eficiência das máquinas de busca. Partindo do pressuposto de que o inglês é a língua mais difundida, principalmente na Internet, seria necessário que o usuário tivesse total domínio desse idioma, para poder compreender os documentos recuperados; mas a dificuldade surge para aqueles que não dominam o idioma, apresentando problemas na leitura dos documentos e ainda na formulação de consultas nos SRI, por vários motivos e, entre eles**:** particularidade de cada língua e contexto semântico dos termos médicos (SCHULZ et al., 2002).

Com base nessa necessidade, surgiu o Projeto *MorphoSaurus*[1] , que já mostrou sua eficiência na recuperação de documentos médicos ao utilizar um tesauro com descritores semânticos abaixo do nível estruturante da palavra (NOGUEIRA et al., 2004).

O Projeto *MorphoSaurus*[2] (acrônimo de ***MORPH****eme* e *the****SAURUS***) utiliza um tesauro, ou seja, um vocabulário controlado, no domínio médico para ser utilizado em um sistema de recuperação de documentos médicos relevantes. O tesauro permite certa coordenação no processo de indexação de conceito para a recuperação de documentos, em um sistema usado para a busca textual potencialmente relevantes de grandes coleções de documentos. A figura 4 apresenta a *home page* do projeto.

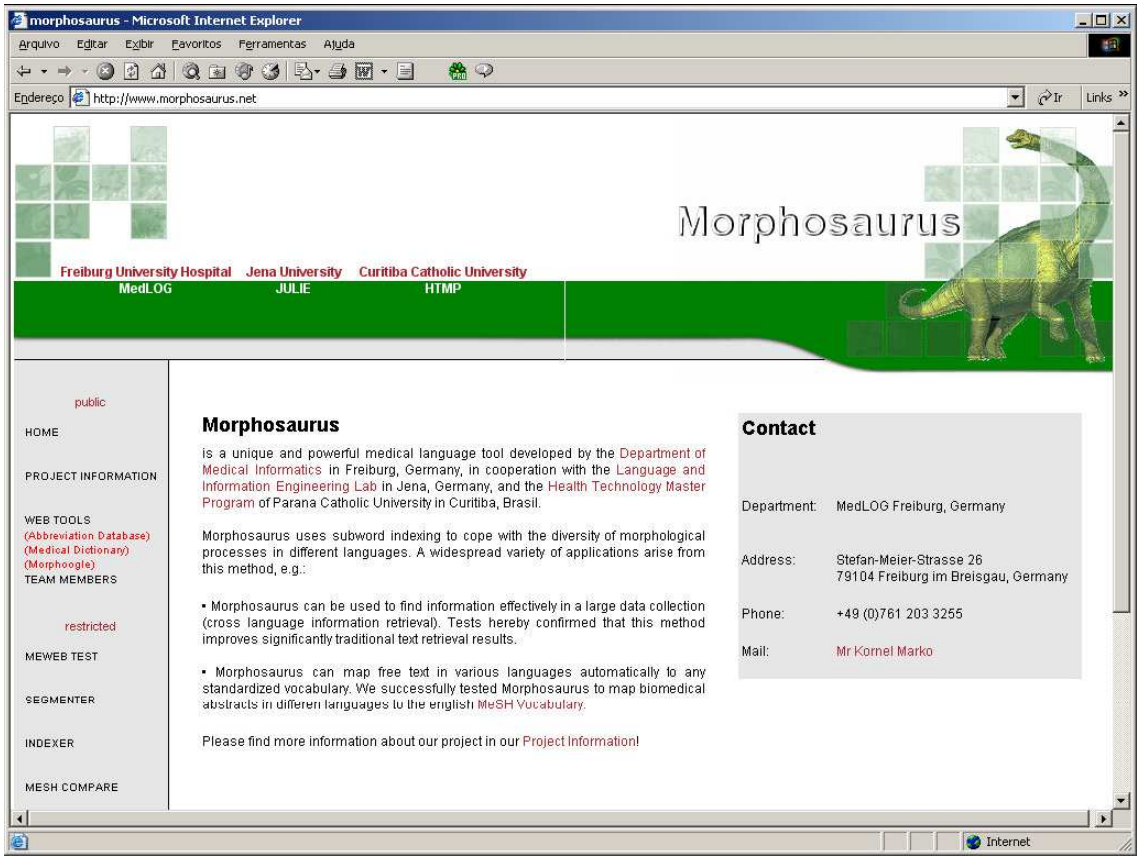


Figura 4 – *Home Page* do projeto *MorphoSaurus*.

A maior particularidade do tesauro utilizado é que suas entidades lexicais correspondem – a maior parte – o que foi definido como "*subwords*". *Subwords*

[1] http://www.morphosaurus.net
[2] Este Sistema desenvolve numa parceria entre a Universidade de Freiburg, Alemanha, e a PUCPR desde 2001.

não são termos e na sua maioria, não são palavras que possam ser encontrados em textos livres. Na sua maioria, *subwords* correspondem a morfemas ou grupos de morfemas. O critério fundamental para delimitação de *subwords* é que elas representem conceitos atômicos relevantes do domínio da medicina, com finalidade de indexação (MARKÓ et al., 2003). As classes de equivalência de *subwords* é a base para a geração dos descritores semânticos, independente de idioma, nomeadamente, *MID* (***M****orphosaurus* ***ID****entifier*).

A idéia principal é normalizar os documentos utilizando a técnica de indexação morfossemântica, para depois serem automaticamente indexados, a fim de melhorar o desempenho do mecanismo de busca (SCHULZ, 2000). Um modelo com base em regras sintáticas para combinações morfológicas em um autômato finito é fundamental para a representação da linguagem artificial no Sistema *MorphoSaurus*.

A indexação morfossemântica traduz os documentos fontes e *queries* (expressões de busca) de uma coleção em uma representação multilíngüe na qual os conteúdos são representados pelas *MIDs*. Esse procedimento é realizado por uma máquina de pré-processamento de documentos que consiste, basicamente, em regras de normalização ortográfica, componentes morfológicos para segmentação de palavras, léxicos de *subwords* para cada língua analisada e um tesauro independente de língua. O sistema *MorphoSaurus* fundamenta-se na suposição de que nem palavras totalmente flexionadas nem as segmentadas heuristicamente constituem o nível apropriado de granularidade para descrição do conteúdo (SCHULZ, 2004).

A grande vantagem do uso de um repositório de *subwords* é o seu tamanho, bem inferior às bases de textos e vocabulários atualmente empregadas assim como sua compatibilidade com termos aglutinados (ex. *hepatopancreático, prebetalipoproteinemia, etc.*), corriqueiros na linguagem médica. Desse modo, uma base de *subwords* requer menos recursos computacionais.

Assim sendo, palavras, amostra de textos ou corpora são repassados para os módulos de pré-processamento do *MorphoSaurus* que os devolve normalizados. Nesse processo, utiliza-se um motor de busca (nesse projeto,

utilizam-se módulos de indexação do motor de busca Lucene[3]) somente para a indexação dos documentos já normalizados e organizar os documentos de acordo com os resultados apontados pelo sistema *MorphoSaurus*.

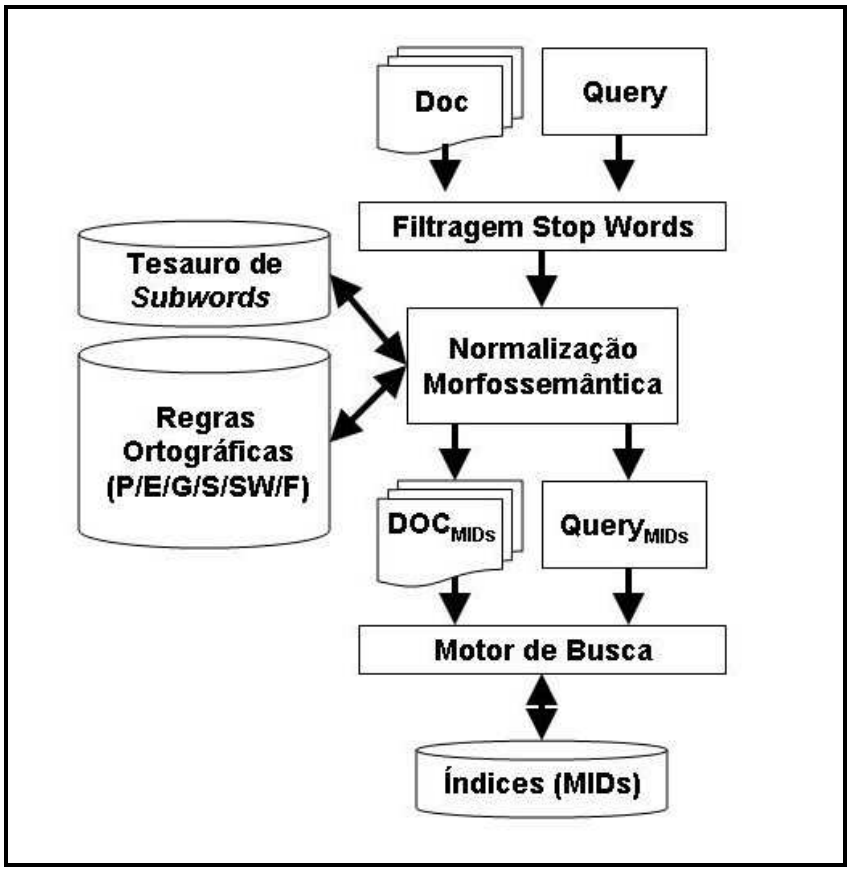


Figura 5 – Representação do processo de normalização morfossemântica do sistema *MorphoSaurus*.

A figura 5 apresenta um esquema do processo de normalização morfossemântica (NOGUEIRA, 2004) executado pelo sistema *MorphoSaurus*; e segue o seguinte fluxo :

(1) documentos multilíngües originais são armazenados e indexados de forma a ficar relacionados com os seus correspondentes normalizados para posterior utilização pelo motor de busca;

(2) antes do processo de normalização morfossemântica os documentos são submetidos a uma filtragem de s*topwords* - conjunto de palavras que não possuem uma representação significativa como artigos, preposições e etc.;

(3) os documentos de cada idioma, com base em regras ortográficas da própria língua e no tesauro de *subwords* são submetidos ao processo de normalização

[3] http://lucene.sourceforge.net/talks/inktomi

morfossemântica e mapeadas para a linguagem artificial representada pelos descritores semânticos através das *MIDs*;

(4) o mesmo processo é aplicados às *queries*;

(5) uma vez normalizados e indexados, todos os termos é possível realizar busca de documentos em línguas que não a nativa.

2.4.2 Indexação

A tabela 1 indica a forma como os documentos são convertidos em descritores semânticos (multilíngüe) através de três passos. O primeiro passo realiza uma normalização ortográfica. Um pré-processador converte todos os caracteres capitalizados para minúsculo e realiza substituições de caracteres específicas para cada língua (as regras estão contempladas no sistema através de um arquivo chamado "*replacement*"), de forma a facilitar a equivalência entre os *tokens* de texto e as entradas do léxico.

Tabela 1 – Processo de normalização morfossemântica para inglês, alemão e português (NOGUEIRA, 2004).

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 primari o	a current**# value# top#** tsh suggest**# diagnos# hypo# thyroid# first#**

O próximo passo trata da segmentação morfológica. O sistema decompõe o texto normalizado ortograficamente em uma seqüência de *tokens* correspondentes às *subwords* no léxico e restos lexicais (não presente no tesauro).

O resultado da segmentação é verificado por um autômato finito conforme figura 6 que rejeita segmentações inválidas. 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. Se o algoritmo de segmentação não detectar uma leitura válida, a palavra original é restituída.

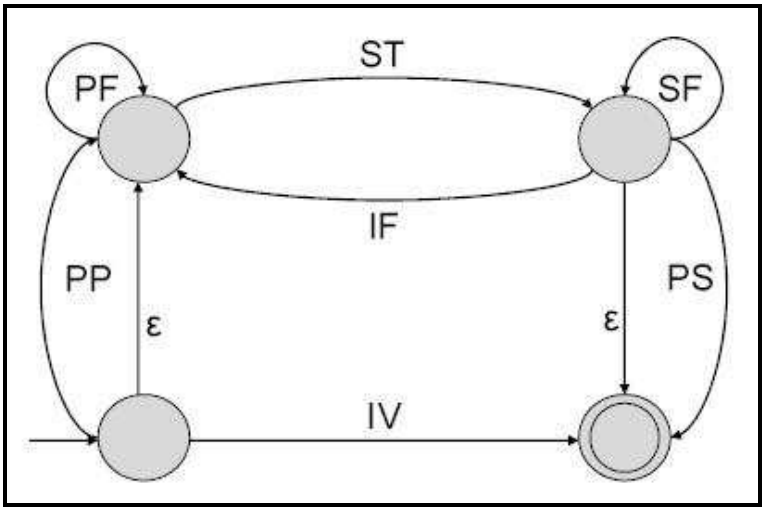


Figura 6: Autômato de estados-finitos para o modelo de *Subword* do Sistema MS com prefixos (PF), prefixos próprios (PP), *stems* (ST), infixos (IF), sufixos (SF), sufixo próprio (PS) e as "*stop entries*" (ε).

No passo final, no processo de normalização semântica, cada *subwords* 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 artificial final.

2.4.3 Montagem 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 quando de forma multilíngüe. Para essa classe de equivalência, estabelece-se um único *MID*.

A criação do tesauro ocorre através de dois tipos de relações para estabelecer vínculos entres as classes de equivalências. Uma relação

sintagmática, pela relação "*has_word_part*" e, pela relação paradigmática, pela relação "*has_sense"*.

A figura 7 contempla um exemplo para dois casos: a relação paradigmática liga um *MID* ambíguo a outros sentidos; isto é, *#head* é ligado aos *MIDs #caput={"cabec-", "kopf", ...}* e *#boss = {"chief", "haeupt"}* pela relação *has_sense*; enquanto a relação sintagmática "*has_word_part*" realiza a ligação de um *MID* aos seus *MIDs* "atomicamente" semânticos; i.e., 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.

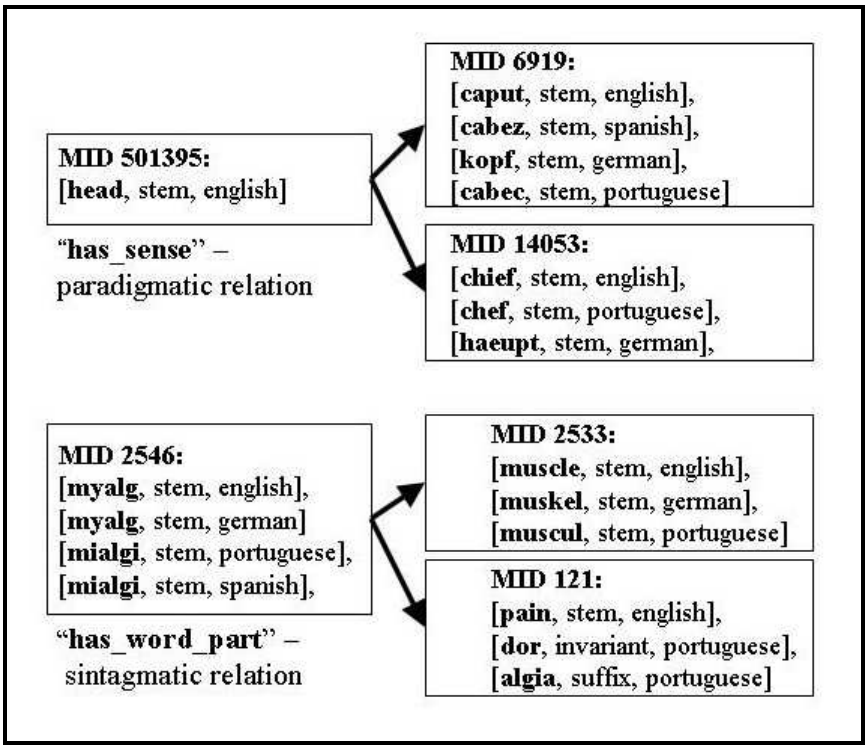


Figura 7: Tipos de Relacionamento semânticos suportados pelo tesauro do MS.

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

(1) Morfemas muito curtos: devido à heurística de segmentação, a segmentação pode levar a interpretações errôneas. Por exemplo:

e_{24} = (*myalg*, ST, #*myalg*, EN, *d*)

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*”

(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 se tem 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 umaboa segmentação, é relacioná-lo na forma seguinte:

$\boldsymbol{R_5}$:= {(#*urinalis,*#*urina*), (#*urinalis, #analis*) } ∈ “*has_word_part*”.

2.4.3.1 Editor de Morfemas (Morphoedit)

A montagem do tesauro é manual, realizada pelos lexicógrafos, onde são utilizadas basicamente três ferramentas para gerenciar este processo de montagem: o gerenciador de tesauro nomeadamente *Morphoedit,* com ferramentas de apoio à decisão e o Segmentador (*MorphoSaurus Segmenter*), para a verificação visual dos resultados da segmentação de expressões regulares.

O *Morphoedit* provê um ambiente multi-usuário baseado na *Web* para criação e manutenção facilitando o trabalho em conjunto dos lexicógrafos em

locais diferentes em níveis nacional e internacional.

Esta ferramenta, conforme ilustrado na figura 8, é aquela que os lexicógrafos utilizam para construção do tesauro, de forma que são realizadas as inclusões, exclusões, alterações e as relações semânticas.

As relações existentes no *MorphoSaurus* são de equivalência (CEq) que é composta por sinônimos e as relações entre as CEq, que aqui são tratados em dois casos: parte de um todo (*has_word_part*) e possíveis acepções (*has_sense*).

As classes de equivalências são compostas por sinônimos de forma a reunir através de relações de equivalência os léxicos das diferentes línguas. Assim, os léxicos, por exemplo, em inglês, "*disease*" e "*illnees*", com os termos alemães, "*krankheit*", espanhol, "*enfermedad*", francês, "*maladie*", sueco, "*sjukdom*" e português, "doença" pertencem a uma única classe de equivalência que representa um sentido.

As funcionalidades da interface, apresentadas na figura 8, são:

(1) em **A,** o usuário tem acesso às ferramentas básicas do ambiente, entre elas: a criação de um novo lexema, a junção entre lexemas existentes, a criação ou remoção de relações entre lexemas, o acesso a ferramentas auxiliares (Mesh, UMLS, *WordStat*), a geração dos arquivos XML e a notificação de *bugs* do sistema;

(2) em **B**, é possível selecionar quais dos idiomas disponíveis serão utilizados na busca;

(3) em **C** ou **F**, lista-se o conjunto de lexemas que satisfazem à condição de busca especificada;

(4) em **D** ou **G**, são listados todos os lexemas que fazem parte de uma mesma Eq Class. Uma Eq Class é um identificador único que agrupa os lexemas, de diferentes idiomas, que tenha mesmo significado semântico. Os *MIDs* são criados a partir de um Eq Class e representa um conceito. De forma que necessita estar selecionado os léxicos em **C** e **F**, para realizar a função *JOIN* em **A**, teremos uma equivalência entre estes léxicos, assim formando uma mesma classe de equivalência;

(5) em **E** ou **H** são listadas as relações da Eq Class selecionada.

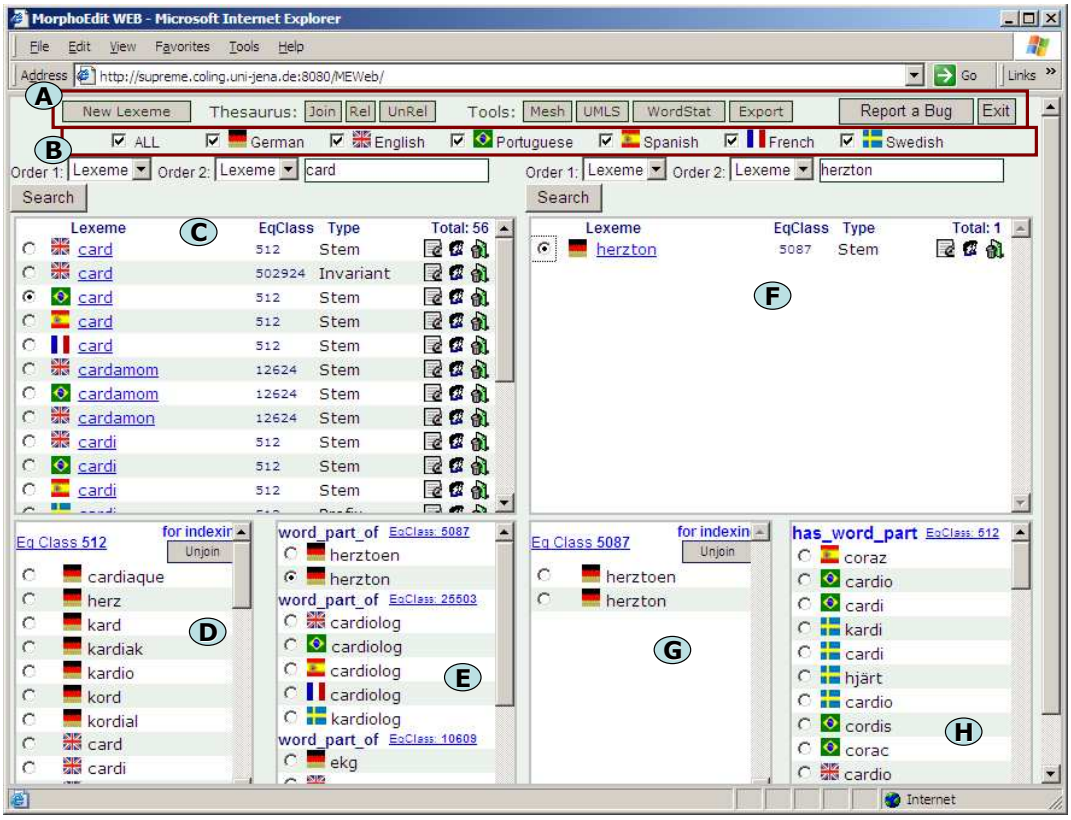


Figura 8 – *Morphoedit*: gerenciador de léxicos.

No *Morphoedit,* podem ser inseridos os lexemas:

(1) *stem*, que são os radicais das palavras, formam a parte principal da palavra, possuindo uma alta carga semântica. Por exemplo, "*gastr*", "*hepat*";

(2) prefixos (*Prefix*): estão a frente de um radical podendo ocorrer mais de uma vez em uma palavra. Por exemplo "*hyper*", "*anti-*";

(3) prefixo próprio *(True Prefix)*: caracteriza prefixos que não podem ser prefixados. Exemplo: "*hemi-*", "*down-*";

(4) sufixos (*Suffix*): são os elementos acrescentados a um radical, ocorrem no depois de um radical, podem ocorrer também após um sufixo, como "*-tomia*", "*ite*";

(5) sufixo próprio (*True Suffix*): caracteriza sufixos que não podem ser sufixados. Por exemplo "*-ação*", "*-ão*";

(6) infixos (*Infix*) são os que ligam dois lexemas, geralmente *stems*, como "*gastr-o-intestinal*".

(7) invariante (*Invariant*): um lexema cadastrado como invariante, faz com que o segmentador, quando encontra tal *string* inteira em um texto, ela assume como única, da forma como foi cadastrada, não segmentando a mesma. Por

exemplo: nomes próprios como “*aspirina*” e acrônimos como “ECG” ou “AIDS”.

O *Morphoedit web* foi desenvolvido na linguagem de programação 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 incluem:

(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).

O *Morphoedit web* possui 3 ferramentas de apoio: *WordStat*, UMLS, MeSH, os quais servem para que os lexicógrafos utilizem-nas para auxiliar na inclusão de novos lexemas no sistema, haja vista a complexidade e o multilíngüismo da terminologia médica. Todas essas ferramentas estão acopladas no editor de morfemas, para facilitar a consulta pelo lexicógrafo.

O *Wordstat* é uma ferramenta implementada que apresenta uma lista de termos em ordem alfabética e por ordem de freqüência de termos utilizados em pesquisa, serve para verificar a atomicidade dos termos, ou seja, o tamanho dos termos que será incluído para um determinado conceito, conforme ilustra a figura 9.

O *Wordstat* dá acesso a uma estatística de distribuição de palavras que foi compilada a partir de *corpora* de referência, extraídas da *Web*

(especialmente do MSD Manual que existe 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 é importante especialmente com fragmentos de palavras curtas de três ou quatro caracteres que às vezes ocorrem em múltiplos contextos.

Um exemplo importante, representado na figura 9, em uma pesquisa pela *substring* "*gen*" (raiz de gene), pode-se notar muito bem nesse resultado que as ocorrências dessa *substring* possui contextos completamente diferentes, o que pode-se justificar a não inclusão de "gen" no léxico.

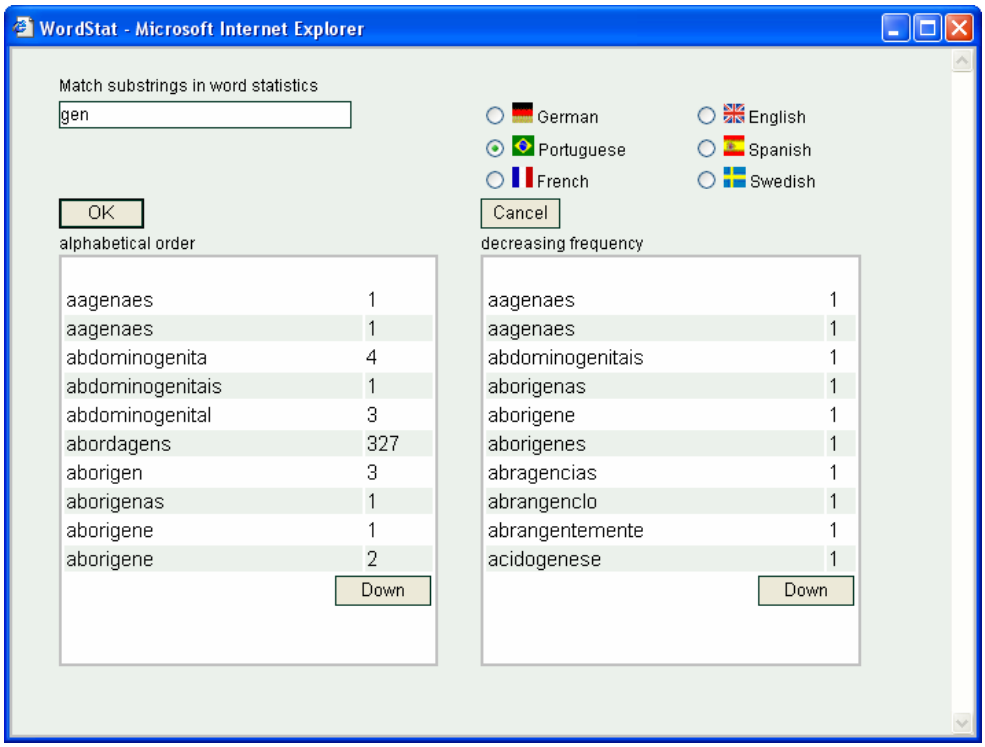


Figura 9 – *Wordstat*: estatística de palavras baseadas em texto do domínio.

O UMLS (*Unified Medical Language System*), vocabulário médico unificado e interligado através de conceitos, sofrem atualizações constantes referentes aos setores da área médica (UMLS, 1994). Essas atualizações são importantes, pois a terminologia médica é vasta e ambígua, denotando um vocabulário atualizado para o entendimento de tal terminologia .

Utiliza-se o UMLS para contemplar as relações de sinonímia entre termos médicos intra e multilingual, além de verificar os termos ambíguos, conforme ilustra a figura 10.

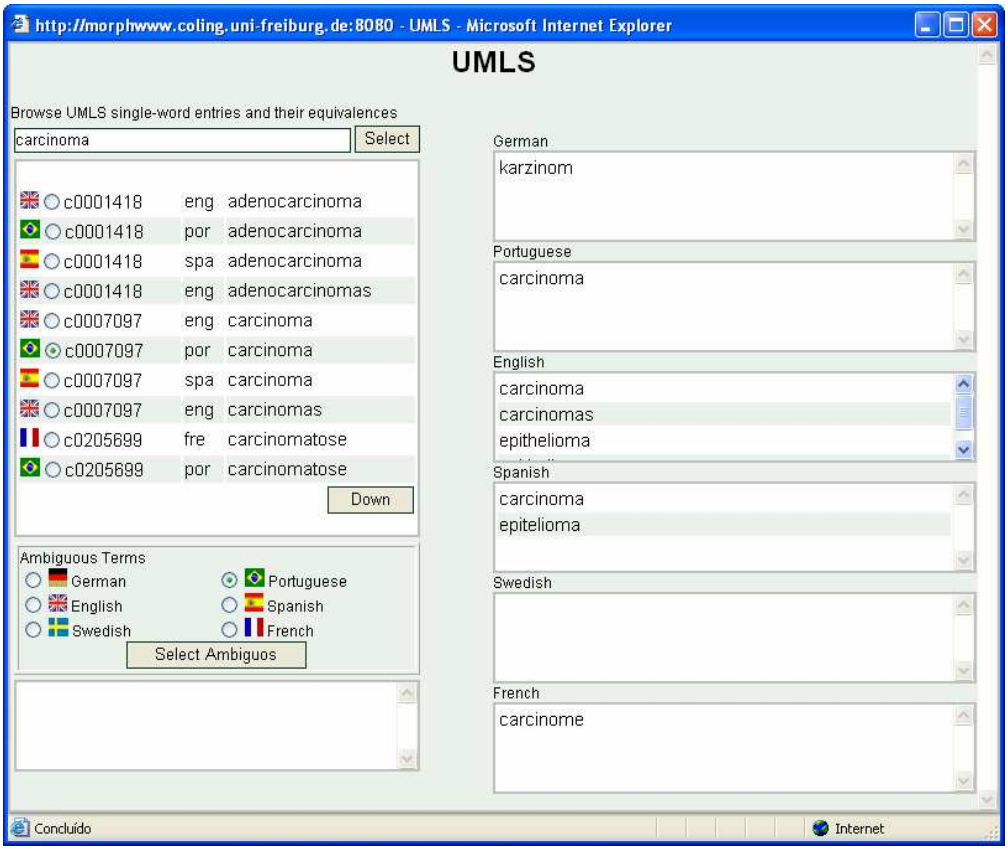


Figura 10 – Ferramenta utilizando o UMLS.

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 têm 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.

O Medical Subject Headings (MeSH) foi criado pela National Library of Medicine 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 (NLM, 2000). 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, assim como categorias mais abrangentes podem recuperar artigos indexados com categorias mais restritas. 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 MeSH constitui uma das fontes do UMLS *metathesaurus* e, por isso, os termos não compostos já vêm na ferramenta auxiliar “UMLS”. Ao contrário deste, a ferramenta MeSH contém os termos completos que normalmente são compostos entre duas e cinco palavras. Desta forma, os lexicógrafos podem ver a ocorrência de uma palavra ou um *substring* em um contexto mais amplo, o que ajuda a detectar acepções adicionais que tem que ser contemplados na construção do léxico, conforme ilustrado na figura 11.

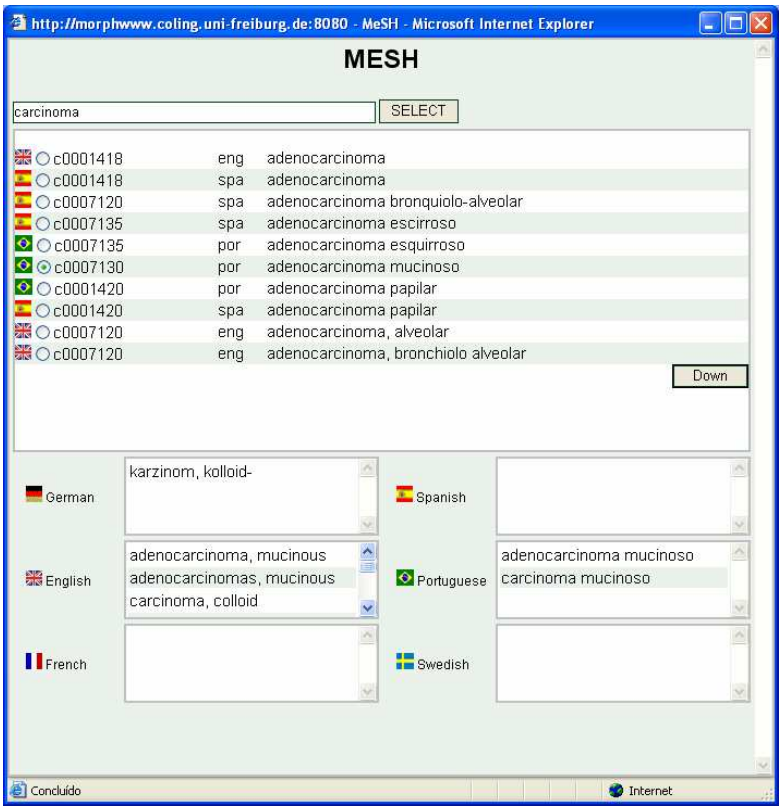


Figura 11 – Ferramenta utilizando o Mesh.

No processo lexicográfico, o uso do segmentador tem um valor heurístico no sentido de verificar o resultado da segmentação de uma palavra ou uma lista de palavras assim como amostras de textos. Ele permite realizar comparações entre textos paralelos em línguas diferentes e dicionários diversos para identificar erros e detectar ambigüidades.

Cada idioma possui seu próprio segmentador devido às regras gramaticais serem diferentes, as regras são inseridas em arquivos distintos. O módulo segmentador pode ser formado por uma palavra ou uma lista de palavras indicando a URL que se encontra, conforme figura 12.

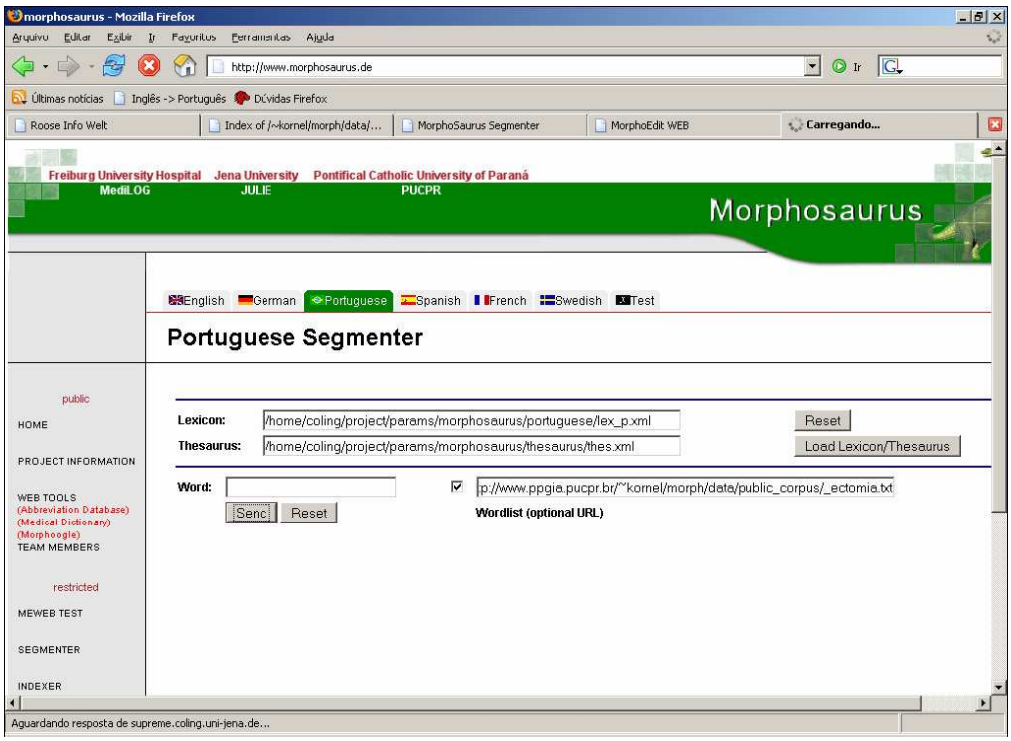
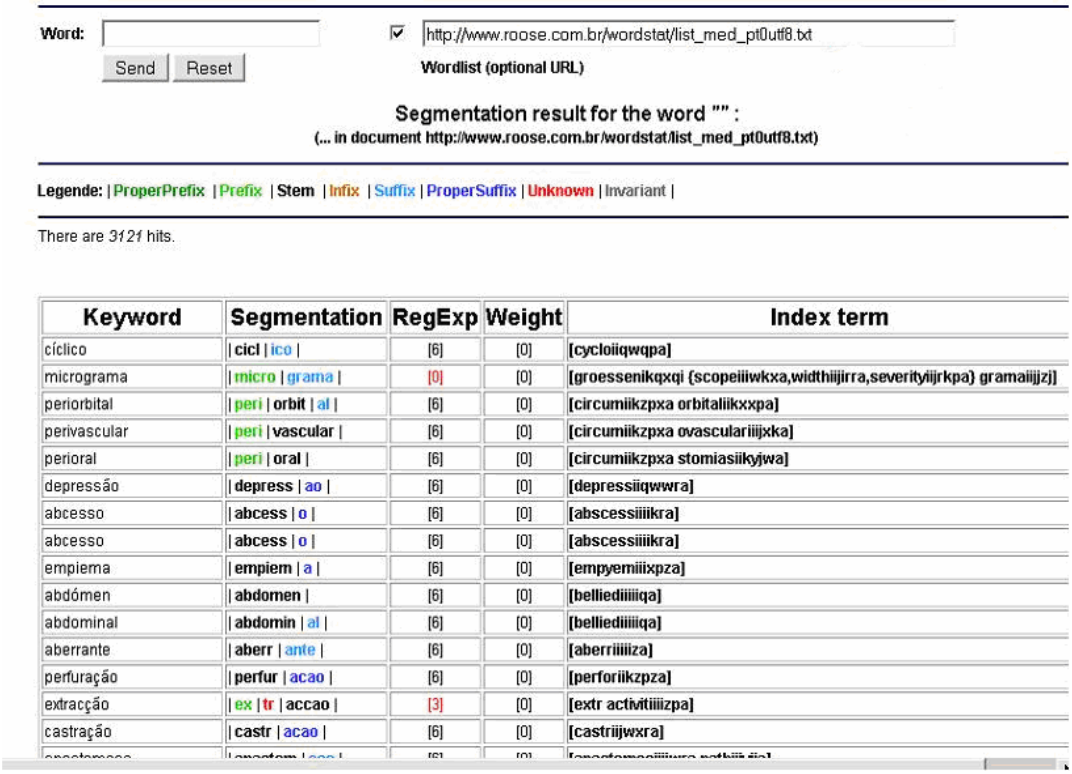


Figura 12 – Módulo segmentador do *MorphoSaurus*.

Após informar os dados para o segmentador, ele gera a partir do texto informador ou lista de palavras, o resultado segmentado e identificado, conforme uma legenda, apresentada na figura 13.



Keyword	Segmentation	RegExp	Weight	Index term
cíclico	|cicl|ico|	[6]	[0]	[cycloiiqwqpa]
micrograma	|micro|grama|	[0]	[0]	[groessenikqxqi {scopeiiiwkxa,widthiijirra,severityiijrkpa} gramaiijzj]
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]	[belliediiiiiqa]
abdominal	|abdomin|al|	[6]	[0]	[belliediiiiiqa]
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 13 – Resultado da segmentação.

2.4.4 Exportação do Tesauro

O XML (*Extensible Markup Language*) é um padrão para publicação, combinação e intercâmbio de documentos multimídia, desenvolvido pelo consórcio W3C (World Wide Web Consortium) (XML, 2001).

A definição da linguagem XML consiste em padrão de marcação com um conjunto de "*tags*", onde há informações estruturadas, ou seja, documentos que contêm estrutura clara e precisa da informação que é armazenada em seu conteúdo (OLIVEIRA, 2002).

O padrão XML foi escolhido por se tratar de uma plataforma aberta de fácil adaptação, além de poder também integrar-se a outros sistemas de busca.

Os lexemas do tesauro são exportados para o padrão XML para cada língua, com o nome "lex_LANG.xml" (onde LANG = PT, EN, GE, SW, SP, FR, etc...) e os relacionamentos também são contemplados no mesmo padrão com o nome de thes.xml. A figura 14 mostra a estrutura interna padrão de um arquivo XML referente ao tesauro e a figura 15 se refere às regras de normalização ortográficas (para ser utilizado com caracteres em ASCII / 7 bits) também definidas em arquivo no padrão XML (replacement.xml) utilizados pelos módulos de segmentação do Sistema *MorphoSaurus*.

O tesauro no padrão XML (thes.XML) gerado, figura 14, possui as seguintes *tags*:

(1) "<lex>" que determina o início e o final de cada lexema;

(2) "<mid>" que representa o conceito de forma multilíngüe - é a linguagem artificial do *MorphoSaurus*;

(3) "<str>" determina qual é a seqüência de caracteres;

(4) "<t>" determina qual é o tipo do lexema, por pedido do projeto *MorphoSaurus* é representada por siglas, ST – Radical, PF – Prefixo, SF – sufixo, IV – invariante, IF – Infixo, PPF – Prefixo próprio, SSF – Sufixo próprio;

(5) "<l>" determina a língua a qual pertence aquele lexema, sendo 1 para alemão, 2 para o inglês, 3 para o português, 4 para o espanhol, 5 para o francês e 6 para o sueco.

O arquivo *replacement* (figura 15) contém as expressões regulares de

pré-processamento para cada idioma.

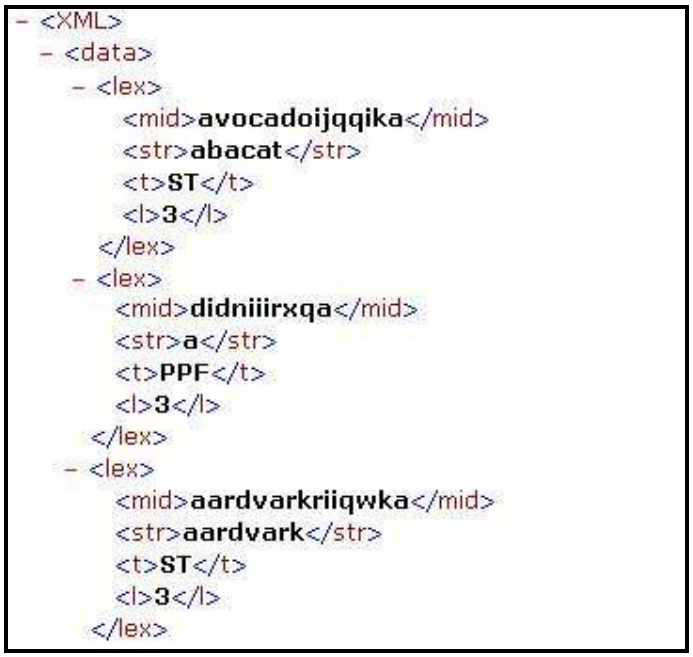


Figura 14 – Representação da Estrutura XML lex.

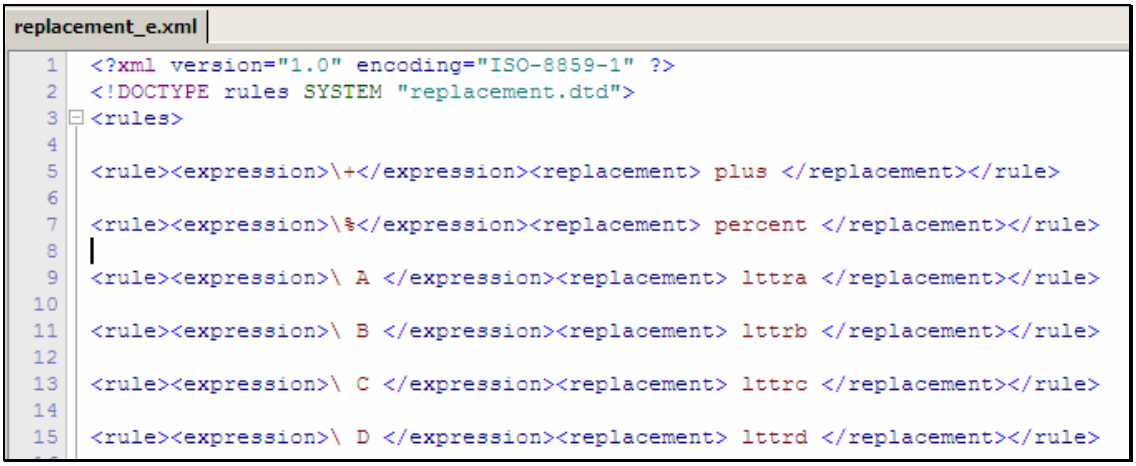


Figura 15 – Representação das regras de substituição na estrutura XML.

CAPÍTULO 3

METODOLOGIA

Nesta pesquisa, propõe-se a criação de um registro de procedimento para identificar os processos anômalos ocorridos no processo de engenharia do léxico/tesauro, no momento de sua criação e manutenção pelos lexicográficos.

A metodologia foi desenvolvida para o sistema *MorphoSaurus*. Porém, o problema subjacente é mais geral e aplica-se em ambientes em que um grupo distribuído de usuários coopera no processo de manutenção de recursos terminológicos ou lexicais.

A metodologia está dividida em 3 partes sendo: coleta de dados, identificação das anomalias e validação, conforme ilustra a figura 16.

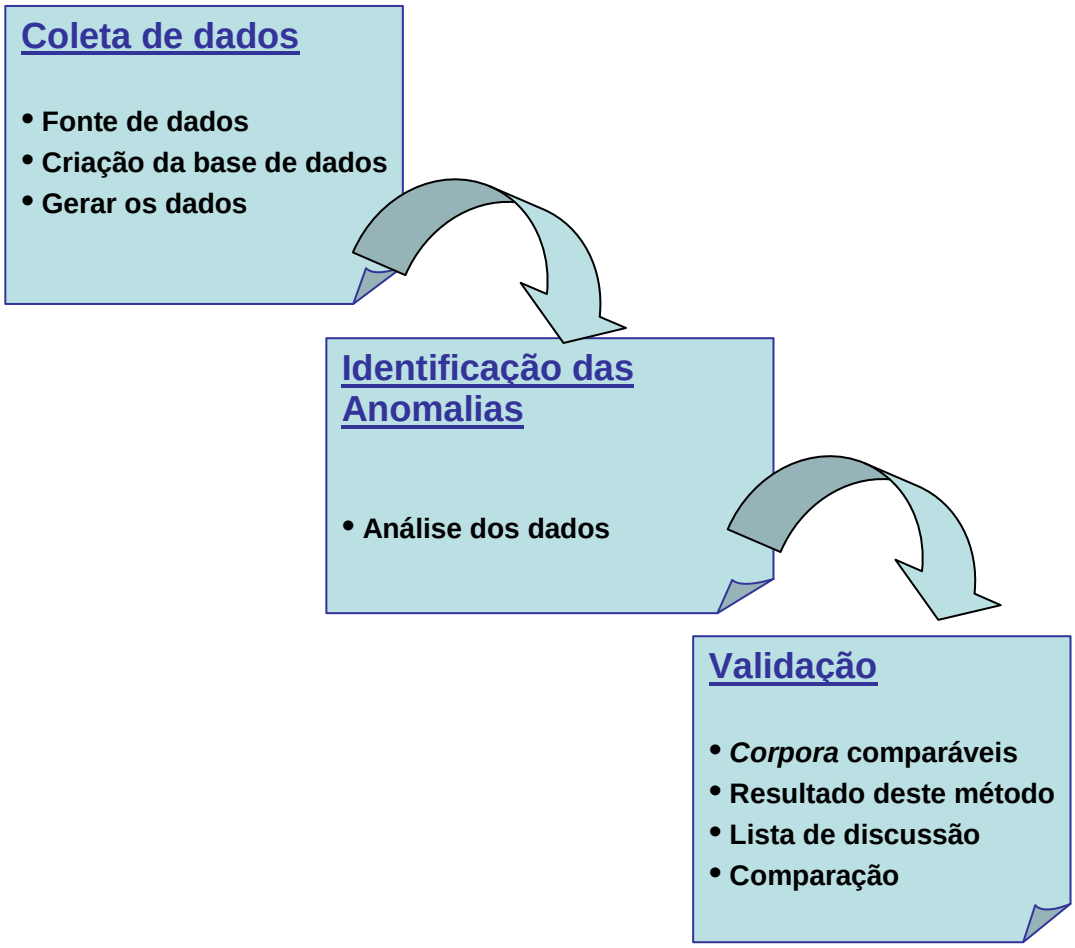


Figura 16 – Representação da metodologia empregada no desenvolvimento da pesquisa.

Os itens da metodologia apresentada na figura 16 são:

Coleta de dados:

(1) fonte de dados: utilizaram-se as cópias de segurança da base do sistema *MorphoSaurus* que permitem reconstruir as ações realizadas por um grupo de lexicógrafos, no período de 14/07/2005 a 30/03/2006. Essas cópias haviam sido arquivadas em intervalos de aproximadamente 3 dias, compondo desta forma 86 bancos de dados restaurados no sistema MySQL5, no formato MEDB_YYYYAAMM, para não haver sobreposição de um pelo outro. O resultado é uma coleção de 86 bancos de dados independentes e distintos por data;

(2) criação de base de dados: criou-se uma base de dados chamada de *log_thes*. Tal base é o local composto por tabelas, onde ficam registrados os dados oriundos da fonte de dados. Esse registro são os dados que serão apresentados neste trabalho;

(3) gerar os dados: para a geração dos dados de testes, criou-se um programa para verificar as alterações ocorridas entre uma base de dados, ou seja, no léxico/tesauros, com relação ao seu anterior e registrá-las em outra base distinta, denominada *log_thes*.

Identificação das anomalias:

Para a geração dos dados foram utilizadas as tabelas do *log_thes*. Por meio desses registros, foram realizadas avaliações para classificar e avaliar os processos anômalos, ocorridos neste período sobre os possíveis problemas de anomalias de relacionamento, tipo, delimitação e permanência.

Validação:

(1) *corpora* comparáveis: foi realizada por parte do grupo de pesquisa uma avaliação no léxico/tesauro, utilizando uma outra abordagem baseada em *corpora* comparáveis;

(2) resultado do método: o resultado da validação com *corpora* comparáveis gerou um índice de possíveis erros dentro do tesauro;

(3) lista de discussão: foi registrada em um fórum de discussão, de modo a serem levantados, principalmente, os tipos de anomalias para posterior análise. Com isso, quando existia um consenso entre todos os lexicógrafos, era registrada a alteração no tesauro;

(4) comparação: para validação, foram cruzados os dados mapeados do registro de procedimento (identificação das anomalias), que possuíam os processos anômalos de cada possível problema, comparadas com as alterações realizadas com base na lista de discussão gerada através de *corpora* comparáveis, que foram registradas no período considerado através de uma lista de discussão, com o histórico da situação anterior e posterior à modificação.

Resumidamente, efetuou-se uma avaliação em dados históricos de um método de controlar a qualidade do léxico ao rastrear o processo lexicográfico durante um período de teste. Paralelamente, implementou-se um protocolo de detecção e eliminação de erros acompanhando o mesmo processo, baseado em uma abordagem empírica e usando *corpora* de teste. A questão científica consiste em elucidar se os dois métodos de controle de qualidade coincidiam em detectar os mesmos erros.

3.1. COLETA DE DADOS

Para atender à necessidade de mapeamento dos procedimentos realizados pela equipe de lexicógrafos, criou-se um base de dados onde são registrados todos os eventos referentes às modificações realizadas no tesauro, seja em nível de classes de equivalências, em nível de relações entre classes relacionamentos ou a nível das propriedades e da eliminação do lexema, de forma a guardar o histórico cronológico dos procedimentos. As informações armazenadas possuem a função de gerar dados estatísticos relacionados aos vários estados de uma classe ou de um lexema; por exemplo, o histórico de suas acepções (*has_sense*), suas composições (*has_word_part*), o histórico de um lexema como sua delimitação morfossintática ou as classes por onde este lexema pode ter migrado.

Para isso, foi criada uma base de dados distinta do tesauro chamada *log_thes*, a qual é composta pelas tabelas de registro de lexemas, registro de equivalência entre lexemas e registro de relações semânticas, sendo que esses registros compõem os dados para o sistema de mensagem.

3.1.1 Registro de Lexemas

Os lexemas são registrados na tabela do banco de dados denominada *Log_BaseLexeme* que é composta pelos sete atributos relacionados na tabela 2. Nesta tabela, todas as alterações que ocorreram nos lexemas são registradas.

Tabela 2 – Log_BaseLexeme registrada no banco.

Atributos	**Descrição**
OPERATION	registra o tipo de operação realizada, que neste caso será efetivado com o *CREATE*, *DELETE* e *EDIT*
ID_LEXEME	o número do lexema no banco
M_STRING	a *string* do lexema
TYPE	o tipo do lexema (*stem*, invariante, prefixo, sufixo, infixo, prefixo próprio e sufixo próprio)
LANG	a língua do lexema
EXAMPLE	registra um exemplo do lexema
EQ_CLASS	registra o número da classe de equivalência
DATETIME	registra a data e hora da operação

Neste caso, foram realizadas as ações entre os bancos na seguinte seqüência:

(1) exclusão (*delete*): verificando dois bancos ao mesmo tempo e retornavam os dados referente a esta tabela de registro de lexema, que estavam na base de dados antigo (BDA) e não estavam na base de dados novo (BDN);

(2) inserção (*insert*): analisando nos dois bancos os dados que não estavam no BDA e estavam no BDN;

(3) edição (*edit*): foram registrados na tabela 2 os dados que estavam no BDA e também estavam no BDN, mas que sofreram alguma alteração.

3.1.2 Registro de Equivalência entre Lexemas

Registra-se na tabela do banco de dados denominada Log_Equi, composta por 6 atributos, que estão relacionados na tabela 3, que cuida do

registro de todas as equivalências entre lexemas (sinônimos).

Tabela 3 – Log_Equi registro de todas as equivalências entre lexemas.

Atributos	Descrição
OPERATION	Registra o tipo de operação realizada: *JOIN* ou *UNJOIN*
EQCLASS1	O identificador do conceito 1
EQCLASS2	O identificador do conceito 2
DATETIME	Registra a data e hora da operação
DETAIL_EQCLASS1	Os lexemas pertinentes a EQCLASS1
DETAIL_EQCLASS2	Os lexemas pertinentes a EQCLASS2

Nesta tabela, foram registradas ações realizadas entre os bancos, e os procedimentos seguiram a seqüência descrita a seguir:

(1) com comparação entre dois bancos de dados, sendo que todos os lexemas que estavam no BDA e que não estão no BDN, e se estavam presentes todos em uma mesma classe do BDN, realizou-se uma junção de termos sinônimos que se denomina de *JOIN*;

(2) os lexemas que não estão nas mesmas classes de equivalência do BDA para o BDN, quando não são todos os lexemas, registrou-se a operação de retirar destes lexemas a classe de equivalência, como resultado desta avaliação tem-se os lexemas na mesma classe de equivalência não são sinônimos, esta operação chamamos de *UNJOIN*. Se todos os lexemas removidos em uma classe de equivalência existentes no BDA estão em outra classe já existente, realizar um *JOIN*, da classe gerada pelo *UNJOIN* e a *eqclass* existente;

(3) os lexemas que não estão nas mesmas classes de equivalência do BDA para o BDN, quando não são todos, registrou-se a operação *UNJOIN*. Mas se alguns foram para uma classe de equivalência e outros para outras classes existentes no BDA, realizar um *JOIN* um a um.

3.1.3 Registro de Relações Semânticas

Na tabela do banco de dados denominada Log_Relations, que é composta por cinco atributos, relacionados na tabela 4, registram-se todas as

relações semânticas dentro do tesauro.

Tabela 4 – Log_Relations registro de todas as relações entre as classes de equivalências.

Atributos	Descrição
EQCLASS1	primeira *EqClass* da relação
ACTION	o tipo de relação entre as classes de equivalência
EQCLASS2	segunda *EqClass* do relacionamento
DATETIME	data e horário da operação
IS_ENABLED	se a relação está ativa ou não

Neste caso, somente serão registradas no ACTION as relações de *has_sense* e *has_word_part*, pois são as duas relações realizadas pelo *Morphosaurus*.

O valor do campo IS_ENABLED evidenciará a validade ou não da relação.

Esta tabela foi criada e registra todas as operações dentro do sistema, não havendo a necessidade de realizar um programa para nela registrar os dados, pois o programa existente já estava registrando tais informações.

3.2. ANÁLISE DE ANOMALIAS

Foram registradas todas as operações relacionadas à manutenção do tesauro, nas tabelas descritas anteriormente na coleta de dados. O analisador de anomalias baseia-se na freqüência das alterações.

A partir dessas tabelas, são registrados os dados que basicamente mapearam quatro tipos de situações freqüentes no tesauro: anomalia de relacionamento, anomalia de tipo, anomalia de delimitação e anomalia de permanência.

3.2.1 Anomalia de Relacionamento

O registro do procedimento deste tipo de problema é proveniente da tabela Log_Relations, que registra as relações semânticas. Assim, foram

registradas todas as operações relacionadas à manutenção do tesauro, ou seja, todos os procedimentos envolvendo a criação, mudança ou eliminação de relacionamentos entre classes (*has_sense* e *has_word_part)*.

A partir disso, foi analisado este possível problema detectando as ocorrências de anomalias de delimitação semântica.

As anomalias, neste caso, eram registradas no momento em que as classes relacionadas foram posteriormente quebradas o relacionamento entre elas e posteriormente refeito, como exemplo desta anomalia pode-se ser verificada na figura 17.

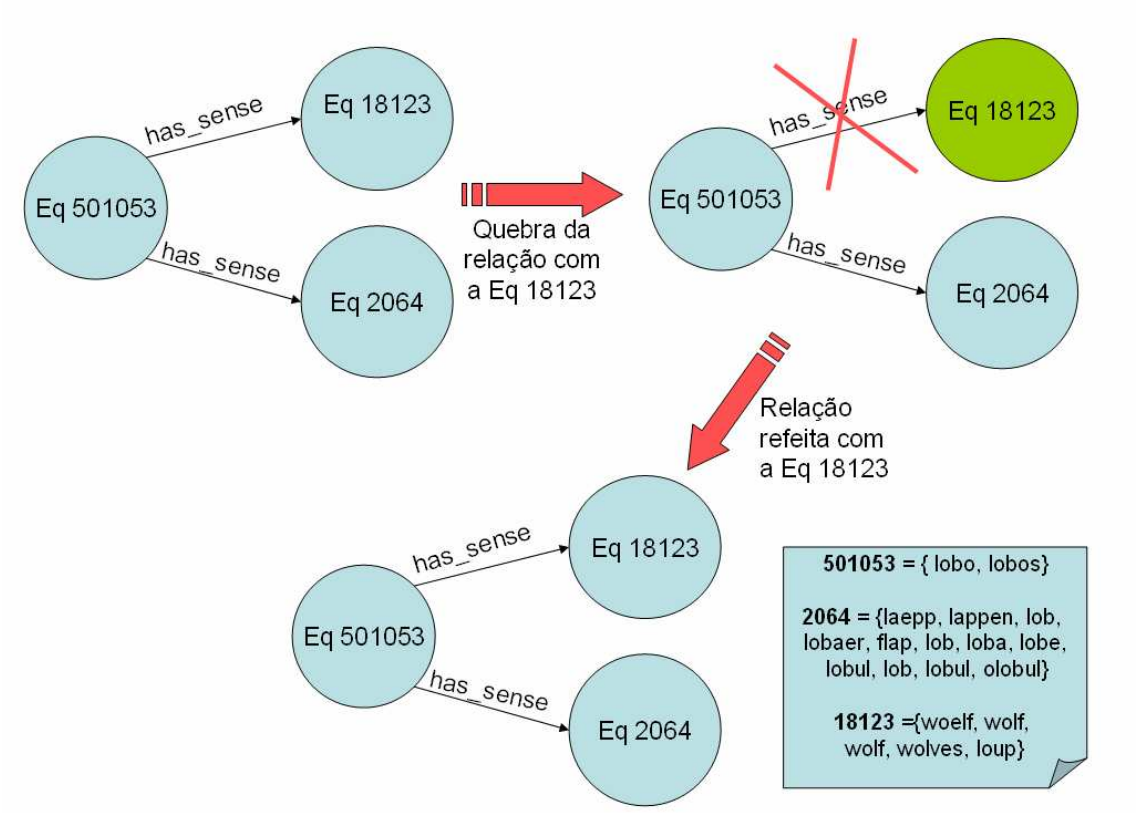


Figura 17 – Exemplo de anomalia de relacionamento.

Um exemplo pode ser visto na figura 17, decorrente da indefinição de escopo (incluir termos biológicos / zoológicos ou não) o primeiro lexicógrafo cadastrou “lobo” como lexema ambíguo, tendo as acepções “*wolf*,...” e “lobe,...” Depois, um outro lexicógrafo eliminou a acepção “*wolf*”, achando que essa homonímia gera ambigüidades desnecessárias em um contexto médico. Depois, outro lexicógrafo restituiu a situação original.

3.2.2 Anomalia de Tipo

Anomalia de tipo corresponde ao problema de inclusão de um lexema em uma classe de equivalência. Esta anomalia baseia-se no processo de equivalência onde os lexemas sinônimos que representam o mesmo conceito são representados por classes.

O registro de procedimentos deste tipo de problema é oriundo da tabela Log_Equi, onde todas as relações de equivalências entre lexemas, como também de suas não equivalências, são registradas.

Neste tipo de problema são registrados, além das classes, os lexemas que as compõem, que são os seus sinônimos.

Foram considerados como anomalias as ocorrências em que os mesmos lexemas foram excluídos e depois novamente incluídos em uma classe de conceito, conforme exemplo da figura 18, onde, em um primeiro momento os termos gastrocnem estavam na Eq 1054, em um segundo momento os lexicógrafos criaram uma nova classe separando o termo e, em um terceiro momento, foi incluído novamente o termo gastrocnem na Eq 1054.

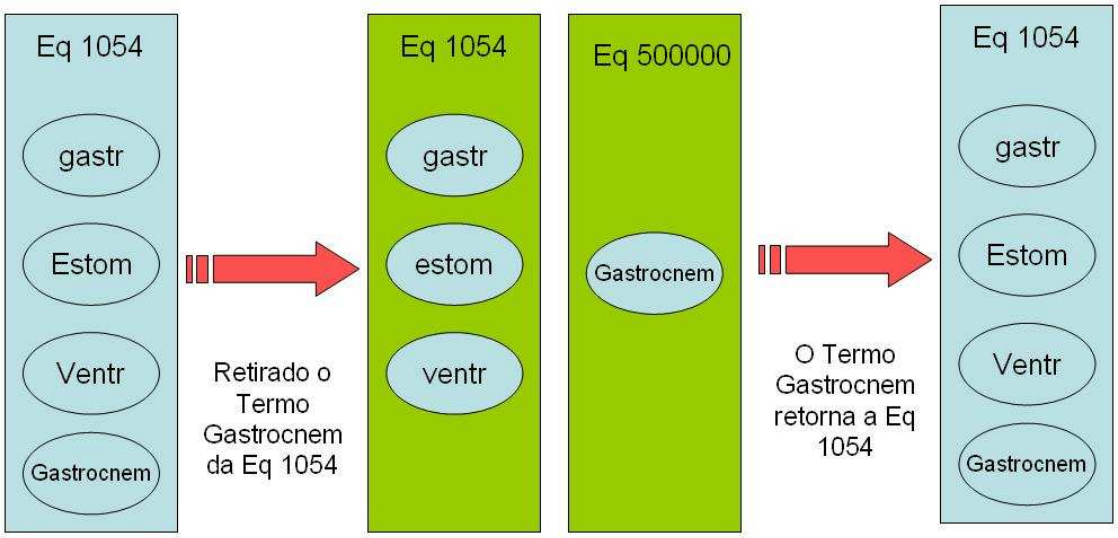


Figura 18 – Exemplo de anomalia de tipo.

Na figura 19, está a forma correta após discussão entre os lexicógrafos. Existem dois músculos do corpo humano cujo o *stem* é "*gastr*": digástrico (dois ventres) e gastrocnemio (músculo da panturrilha), e não existe a relação entre essas classes para não ter confusão entre os órgãos, resolveu-se separar. O

que fica claro, nesse exemplo, é a limitação imposta pela granularidade das entidades lexicais, no sentido que as raízes “*ventr*” e “*gastr*” são usadas igualmente para denotar o ventre do corpo e o “ventre” de um músculo.

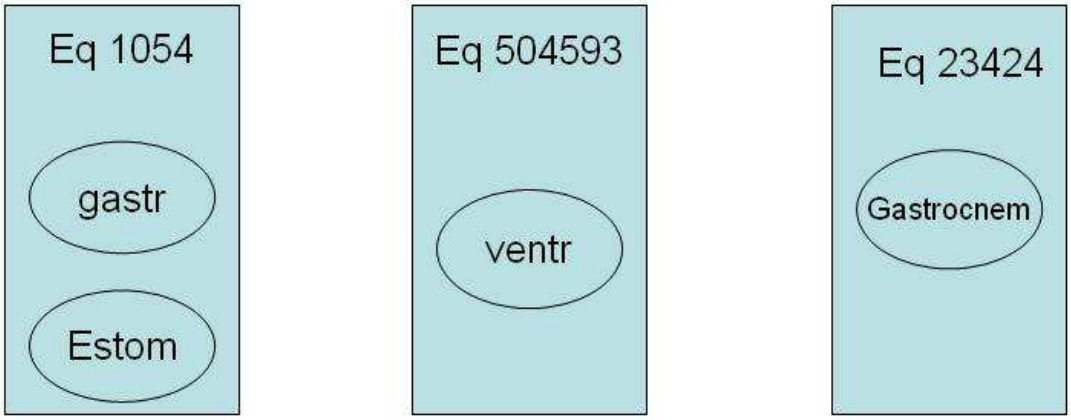


Figura 19 – Exemplo correto da delimitação conceitual.

3.2.3 Anomalia de Delimitação

A delimitação de lexemas, isto é, *strings* que correspondem exatamente a radicais, prefixos, sufixo e afixos, é analisada a partir da tabela Log_BaseLexeme, onde são registradas todas as alterações ocorridas no lexema referente ao seu tamanho. Neste caso, são analisadas somente as alterações cujo campo *OPERATION* desta tabela possua o valor *EDIT*.

Neste tipo de problema, foram somente registradas as alterações realizadas no campo *M_STRING* da tabela, que corresponde ao lexema propriamente dito, cujos campos restantes desta tabela permaneceram estáticos.

Como foi registrada cada alteração, considera-se como anomalia os lexemas modificados que retornaram a qualquer estado anterior à sua alteração, independente do número de alterações, conforme ilustrado na figura 20; onde, em um primeiro momento, o termo *oglob* com suas características está registrado no sistema. Em um segundo momento, houve alteração em seu tamanho, alterando somente o seu tamanho para *glob* e, em um terceiro momento, foi recomposta a sua forma original, sendo que neste processo foi

alterado somente o tamanho; as outras características permaneceram as mesmas.

Figura 20 - Exemplo de anomalia de delimitação.

No exemplo da figura 19, mostram-se os dois tipos de problemas que ocorrem na delimitação do tamanho de um lexema: o sintático e o semântico. O exemplo do sintático é *oglob*, onde foi verificado que é importante colocar *oglob*, pois o segmentador necessita do *oglob* para segmentar corretamente, forma registrado no sistema *glob* e *oglob* como sinônimos e um exemplo dele é globulina e hemoglobina.

Já o semântico, também representado na figura 21, ocorreu com *carcin* e *carcinoma*, onde um lexicógrafo deixava *carcin* e o outro alterava para carcinoma, o problema que o sufixo *OMA* representa um tumor, mas como maioria benigno e já carcinoma são todos malignos.

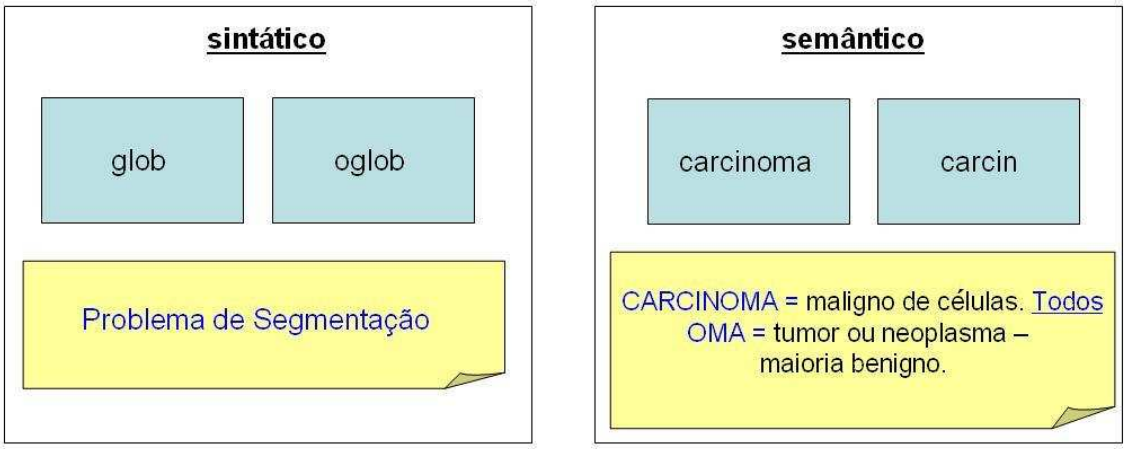


Figura 21 – Tipos de problemas na delimitação do lexema.

3.2.4 Anomalia de Permanência

A relevância lexical foi registrada, neste caso, na mesma tabela da delimitação do tamanho do lexema, a tabela Log_BaseLexeme, onde são registradas todas as alterações ocorridas no lexema, mas somente analisadas quando o campo OPERATION desta tabela é *INSERT* e *DELETE*.

Neste caso, foram registradas como anomalias aqueles casos onde um lexema foi eliminado e depois inserido novamente. Considerava-se somente aqueles cujos dados desses campos eram iguais, conforme exemplo da figura 22, onde *pneumon* que pertence a Eq 13839 que é sinônimos de *pulm* e *pulmon*, significando pulmão. Pois *pneumon+ite* é inflamação nos pulmões, o *pneumon* é pulmão e o sufixo *ite* é inflamação.

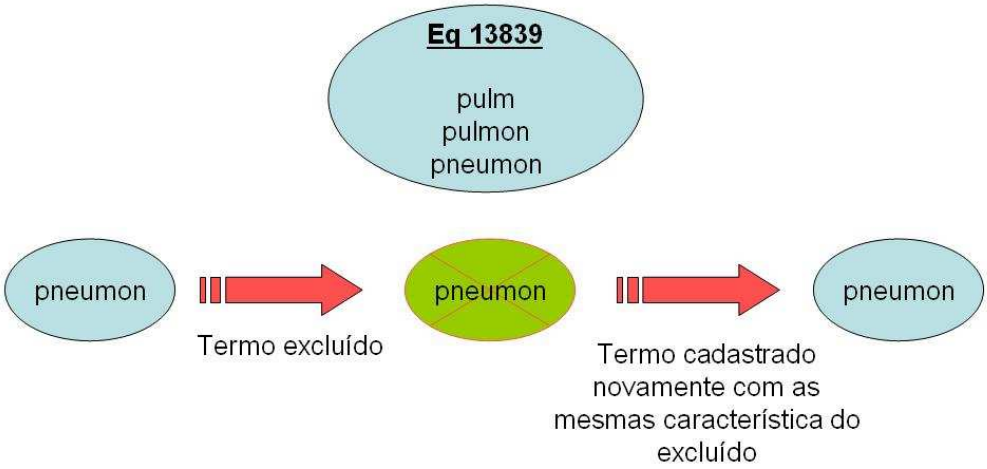


Figura 22 - Exemplo de anomalia de permanência.

3.3. VALIDAÇÃO

No período avaliado, gerou-se uma lista com possíveis problemas no tesauro, por meio de uma outra abordagem, baseada na análise de *corpora* comparáveis. Ao invés dos chamados *corpora* paralelos (que são textos literalmente traduzidos), *corpora* comparáveis são amostras de textos de pares de idiomas ou entre vários idiomas; textos que possuem a mesma função

comunicativa independente do seu idioma (LAFFLING, 1992) e, em conseqüência, cobrem o mesmo espaço conceitual. Daí, decorre a hipótese que motivou o uso de *corpora* comparáveis. Utilizando o *MorphoSaurus* para indexar esses textos, supunha-se que a distribuição de *MIDs* era similar independente do idioma.

A partir dessa hipótese, aplicaram-se um *score*, que indica as ocorrências de *MIDs* numa determinada língua em relação à outra. Foram normalizados os textos e gerada lista para correções e divididas com quantidade equivalente para os lexicógrafos, nas tabelas 5 e 6 encontram-se alguns resultados desta lista, por exemplo, na tabela 5, onde foi percebido a partir deste *score* (maiores índices), que a freqüência de uma palavra normalizada no inglês ocorria em maior quantidade que no texto em português, quando deveriam ser bem próximas ou até mesmo iguais.

O índice (score - S) foi parametrizado, conforme equações (4) e (5) sendo normalizado para que fique com um valor entre 0 e 1. Pressupôs-se que as MIDs próximas da unidade indicam uma maior probabilidade de estar com algum tipo de problema.

$$S = \frac{2S_d + S_a}{3} \qquad \text{(equação 3)}$$

$$S_d = \frac{|f1 - f2|}{|f1 + f2|} \qquad \text{(equação 4)}$$

$$S_a = \frac{fx}{(fx1 + fx2)_{\max}} \qquad \text{(equação 5)}$$

Onde:

f1 é a freqüência da ocorrência de uma *MID* no *corpus* 1(ex. Inglês);

f2 é a freqüência da ocorrência de uma *MID* no *corpus* 2 (ex. Português);

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

Sd expressa um índice com base na diferença de ocorrência de uma *MID* em um *corpus* normalizado em relação a outro; que também pode ser entendida como a ocorrência de um conceito num *corpus* com relação a outro;

Sa relaciona o valor relativo da ocorrência de uma *MID* com relação ao

maior índice de ocorrência em ambas as listas;

S é o índice final com o objetivo de mostrar indícios problemas no tesauro, normalizado entre 0 e 1.

Tabela 5 – Lista de freqüência de *MIDs* entre português e inglês: comparação multilíngüe de ocorrência em *corpora* comparáveis. Essa lista é ordenada pelo grau de disparidade.

MID	MIDCod	f1	f2	Sa	Sd	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 6 – Lista de freqüência de *MIDs* entre alemão e inglês: comparação multilíngüe de ocorrência em *corpora* comparáveis.

MID	MIDCod	f1	f2	Sa	Sd	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
likelihoodrijjzwa	501196	628	0	0,0108	1,0000	0,6703
usefulipxzxia	037970	591	0	0,0102	1,0000	0,6701
auspraegrikwkr	502625	0	555	0,0096	1,0000	0,6699
vorliegriprqi	503540	0	539	0,0093	1,0000	0,6698
overipjqqka	031442	782	1	0,0135	0,9974	0,6695
unusualikprwya	023568	463	0	0,0080	1,0000	0,6693

Os lexicógrafos criaram uma lista de discussão onde foram revistos os maiores *scores* da abordagem de *corpora* comparáveis; nesta lista de discussão foram alterados e registrados os problemas encontrados com o histórico da situação anterior e posterior à modificação, tendo um consenso

entre os lexicógrafos, todas as alterações foram baseadas em um protocolo feito pelos lexicógrafos, representadas na tabela 7, em que:

O *MIDcompare por-eng-pat.lst*: nome do arquivo da lista, na string do seu nome mostra os idiomas que foi relacionado nessa lista.

(1) *Current status in list*: apresenta como está colocado na lista que está exemplificado nas tabelas 5 e 6;

(2) *Current status in tesauro (lexicon)*: este item mostra como está cadastrado o léxico e classes naquele momento;

(3) *Problem description*: o tipo de problema que o lexicógrafo que analisou achou naquele item;

(4) *Solution*: a solução proposta pelo lexicógrafo que analisou aquele item da lista.

(5) *Documentation in Comment field of Eq class*: o que está no campo do comentário da classe.

(6) *Neighborhood*: a ligação da classe, com que tipo de termos ou classe ela está ligada.

(7) *Open questions/todo* – abertura para questões do lexicógrafo que está verificando aquele item da lista.

Tabela 7 – Protocolo da equipe lexicográfica para registro de discussão sobre *MIDs* e alterações realizadas no léxico/tesauro (ANDRADE, 2006).

MIDcompare por-eng-pat.lst
1. Current status in list (Situação atual na lista)
2. Current status in tesauro (lexicon) (Situação atual no tesauro (léxico))
3. Problem description (Descrição do problema)
Kind of problem (Tipo do problema)
4.Solution (Solução)
Reasons (Motivos)
5. Documentation in Comment field of Eq class (Documentação em matéria de comentar Eq classe)
6. Neighborhood (Ligação termos e classes)
7. Open questions/todo (Perguntas Abetas)

Desse modo, para validação deste trabalho foi analisadas essa lista de discussão, dividindo os problemas encontrados pelos lexicógrafos, e avaliados caso a caso, cruzando tais dados com as anomalias encontradas.

CAPÍTULO 4

RESULTADOS

Os resultados obtidos são apresentados neste capítulo com a análise das anomalias cometidas pelos lexicógrafos no tesauro. A identificação de anomalias é usada como técnica capaz de auxiliar na criação e manutenção do mesmo. O capítulo inicia-se com as observações na lista de discussão categorizando os problemas registrados pelos lexicógrafos. Em seguida, são analisados os problemas das quatro classes de anomalias, respectivamente, relacionamento, tipo, delimitação e permanência; comparando-os com a lista de discussão analisada.

4.1. LISTA DE DISCUSSÃO

Cabe ressaltar que os tópicos discutidos nessas listas decorreram do *ranking* de disparidades de distribuição conforme descrito, e analisadas nas línguas inglesa, portuguesa e alemã. Pela lista de discussão, foram observados os problemas e qualificados um a um conforme seu tipo, pelos lexicógrafos. Esses dados foram retirados da lista e categorizados a partir do protocolo da lista de discussão, que ocorreu do Inglês/Português e Inglês/Alemão, e o total geral somando-se as línguas Inglês/Alemão/Português, conforme apresentada na tabela 8.

Não entraram como dados nesse trabalho os tipos qualificados pelos lexicógrafos como “aparentemente classe sem problema”. São classes que foram analisadas pelos lexicógrafos e não foi encontrado nenhum problema aparente.

As relações das três línguas registradas pela lista de discussão apresentaram 325 problemas.

Tabela 8 – Total de problemas da lista de discussão separados por língua. N representa anomalias / total de *MIDs* analisadas.

Problema	**Inglês/ Português**		**Inglês/ Alemão**		**Inglês/ Alemão/ Português**	
	N	**(%)**	**N**	**(%)**	**N**	**(%)**
Sem relacionamentos	26/136	19,12	60/189	31,75	86/325	26,47
Falta de lexema ou classe	36/136	26,47	44/189	23,28	80/325	24,62
Mesmo conceito em duas classes (diferentes)	27/136	19,85	43/189	22,75	70/325	21,54
Dois conceitos (diferentes) na mesma classe	3/136	2,20	8/189	4,23	11/325	3,38
Termo específico do idioma	1/136	0,74	10/189	5,29	11/325	3,38
Problemas ortográficos	9/136	6,62	7/189	3,70	16/325	4,92
Indexação (relevante/ *stop word*)	22/136	16,18	9/189	4,76	31/325	9,54
Delimitação Sintática	9/136	6,62	7/189	3,70	16/325	4,92
Erro de Segmentação	3/136	2,20	1/189	0,54	4/325	1,23

4.2. ANOMALIA DE RELACIONAMENTO

O problema da delimitação semântica foi categorizado pelo registro de anomalias, no qual foi analisado. Esse caso de anomalia foi detectado em 146 ocorrências.

Para obter o registro dessas 146 anomalias, os lexicógrafos modificaram um total de 3.107 relações no tesauro no período da pesquisa, sendo que em cada 21,28 relações efetuadas no tesauro, detectou-se um problema de delimitação semântica, realizado pelo lexicógrafo, representando 4,7% das transações realizadas nas relações.

As 146 anomalias encontradas estão divididas em:

(1) anomalias não debatidas: anomalias que não foram debatidas na lista de discussão;

(2) anomalias debatidas: correspondendo a uma anomalia encontrada na lista de discussão;

(3) anomalias debatidas repetidas: cuja anomalia foi apresentada mais de uma

vez, representando as classes alteradas várias vezes, tendo o mesmo tipo de alteração todas as vezes;

(4) anomalias não debatidas repetidas: anomalias apresentadas mais de uma vez, mas encontra-se na classe que não foi debatida na lista de discussão.

As anomalias não debatidas e as não debatidas repetidas não foram classificadas conforme a língua por não estarem na lista de discussão, mas comtemplam: 36 anomalias não debatidas, representando 24,66% e 50 anomalias não debatidas repetidas, correspondendo a 34,25%.

As anomalias foram identificadas e apresentadas na tabela 9, a qual apresenta as anomalias que foram repetidas várias vezes, de forma a haver uma duplicação do mesmo erro.

Tabela 9 – Categoria de mensagens representadas na lista para anomalia de relacionamento. N representa anomalias / total de anomalias de relacionamento.

Problemas	**Inglês/Português**		**Inglês /Alemão**	
	N	**(%)**	**N**	**(%)**
Anomalias debatidas	6/146	4,11	11/146	7,53
Anomalias debatidas repetidas	12/146	8,22	31/146	21,23

Retirando as anomalias repetidas, pois ocorreram várias vezes para o mesmo problema, obtiveram-se 76 anomalias distintas, sendo que dessas foram debatidas na lista de discussão 27 anomalias apresentadas na tabela 10.

Tabela 10 – Categoria de mensagens não repetidas na anomalia de relacionamento. N representa anomalias de relacionamento / total de anomalias de relacionamento alteradas.

Problemas	**Inglês/Português**		**Inglês /Alemão**	
	N	**(%)**	**N**	**(%)**
Anomalias debatidas	6/76	7,89	11/76	14,47
Anomalias debatidas repetidas	3/76	3,95	7/76	9,21

Dentre essas alterações e que foram consideradas como anomalias, houve classes que foram freqüentemente relacionadas e, posteriormente, quebradas várias vezes, conforme apresentado na tabela 11, mostrando a quantidade que uma classe ocorreu, este tipo de situação que é composta pelas anomalias debatidas repetidas e anomalias não debatidas repetidas.

Tabela 11 – Freqüência de alterações da anomalia de relacionamento.

Número de relações e retirada de relação	Ocorrências			Total ocorrência de anomalias
	Inglês/ Português	Inglês/ Alemão	Anomalias não debatidas repetida	
2	-	1	3	8
4	3	3	8	56
5	-	2	-	10
6	-	-	2	12
7	-	1	-	7
Total	3	7	13	93

Nas anomalias com repetição ocorreram 93 anomalias, para 23 conjuntos de classes alteradas, considerando conjunto de classes, duas classes que foram relacionadas.

Esse tipo de anomalia foi relacionada e comparada com os problemas da lista de discussão. No gráfico 1, apresentam-se os dados referentes ao idioma Inglês/Português e, no gráfico 2, os dados de Inglês/Alemão.

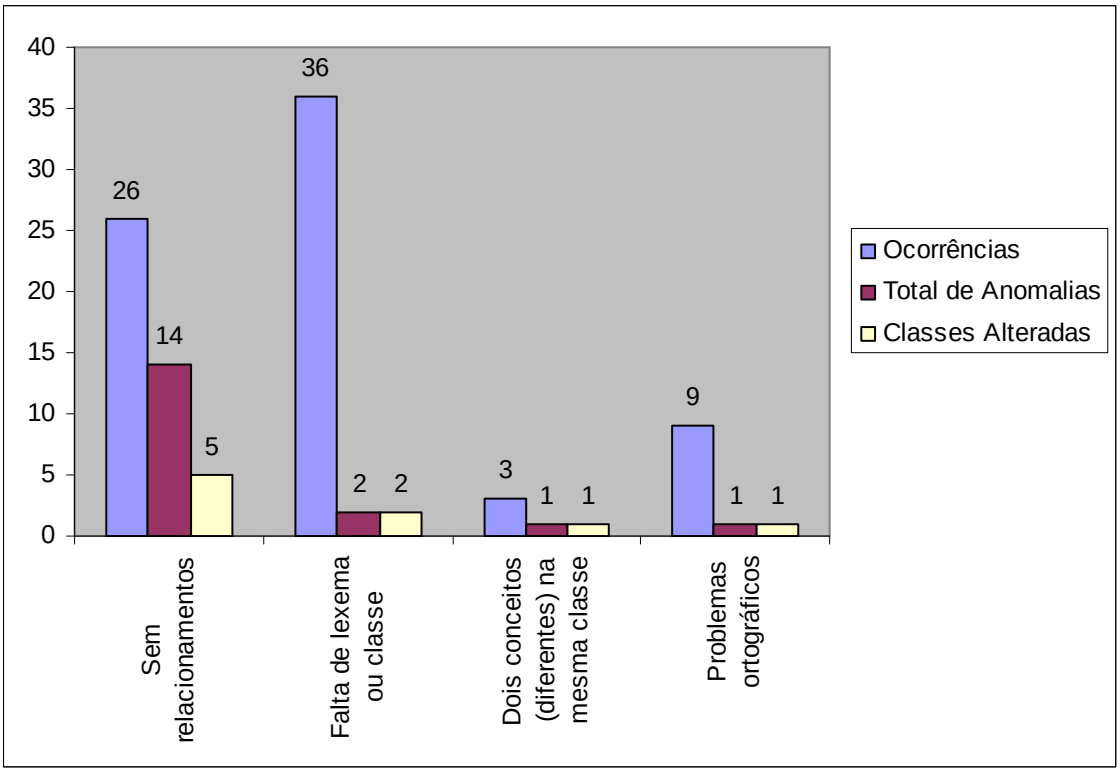


Gráfico 1 – Comparativo na lista de discussão com anomalia de relacionamento em Inglês/Português.

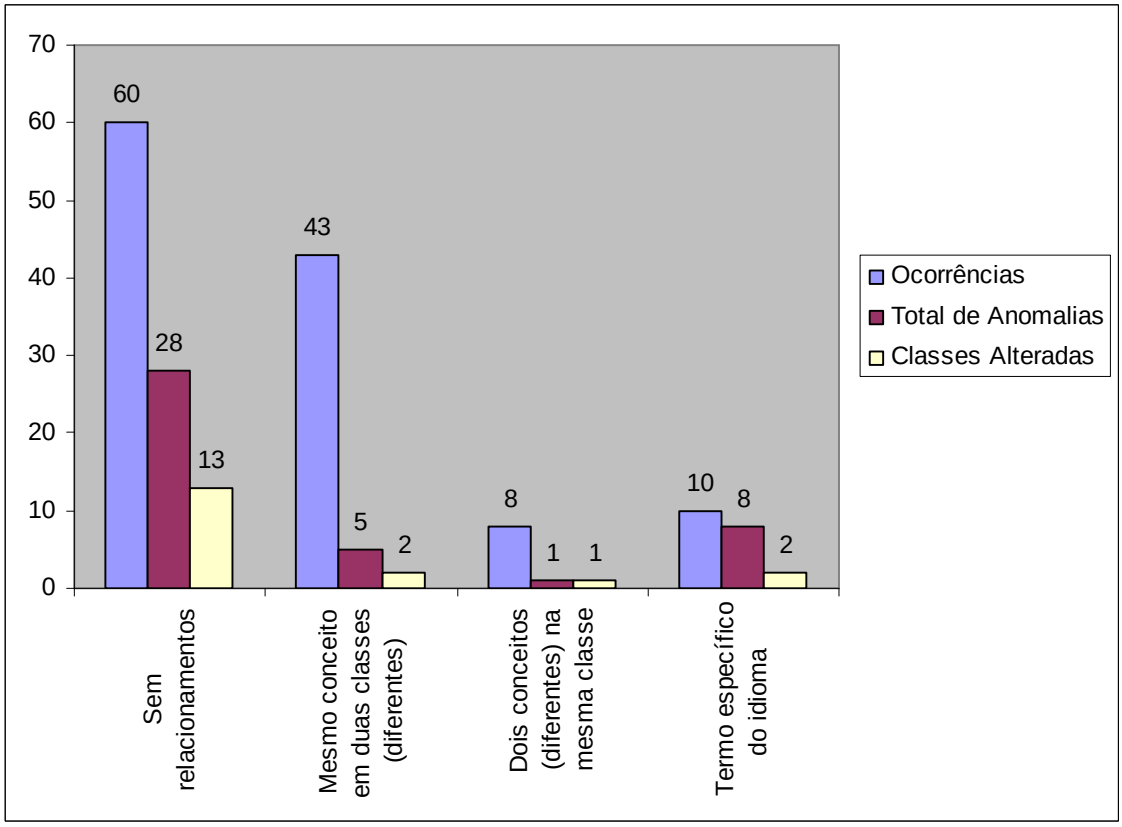


Gráfico 2 – Comparativo entre anomalias de relacionamento em Inglês/Alemão, encontradas na lista de discussão.

4.3. ANOMALIA DE DELIMITAÇÃO

Não foi encontrada nenhuma anomalia deste tipo no período da pesquisa, tendo ocorrido um total de 135 modificações realizadas pelos lexicógrafos, que não se confirmaram como anômalas.

4.4. ANOMALIA DE PERMANÊNCIA

Quanto à relevância lexical, foram inseridos pelos lexicógrafos 983 lexemas e excluídos 133 lexemas. Dentre estes, foram encontrados 5 anomalias com esse tipo de problema, não havendo nenhuma anomalia repetida, das quais, 4 são discutidas e somente 1 não estava relacionada na

lista de discussão.

Relacionando os problemas da lista de discussão e esta anomalia, encontrou-se a relação de somente um problema descrito da lista, com os 4 casos de anomalias debatidas, conforme ilustrado na tabela 12.

Tabela 12 – Comparativo da lista de discussão com anomalia de permanência.

Problemas	Inglês/Português	Inglês /Alemão
Falta de lexema ou classe	1	3

4.5. ANOMALIA DE TIPO

Na anomalia de tipo foram realizadas 2.488 equivalências de classes pelos lexicógrafos, dentre os quais foram identificadas 18 anomalias, não ocorrendo nenhuma repetição.

Dessas 18 anomalias, todas foram encontradas na lista de discussão, conforme apresentado na tabela 13.

Tabela 13 – Quantidade de anomalia de tipo comparando com a lista de discussão.

Problemas	Inglês/Português	Inglês /Alemão
Mesmo conceito em duas classes	3	5
Sem relacionamentos	4	6

4.6. QUADRO GERAL DAS ANOMALIAS

A distribuição total das anomalias, separando por par de línguas e o total geral, conforme a lista de discussão, está representada na tabela 14.

Tabela 14 – Quantidade total de anomalia comparando com a lista de discussão. L representa o total encontrado na lista de discussão e N representa o resultado nesta abordagem.

Problema	**Inglês/ Português**		**Inglês/ Alemão**		**Inglês/ Alemão/ Português**	
	L	**N**	**L**	**N**	**L**	**N**
Sem relacionamentos	26	9	60	19	86	28
Falta de lexema ou classe	36	3	44	3	80	6
Mesmo conceito em duas classes (diferentes)	27	3	43	7	70	10
Dois conceitos (diferentes) na mesma classe	3	1	8	1	11	2
Termo específico do idioma	1	0	10	2	11	2
Problemas ortográficos	9	1	7	0	16	1
Indexação (relevante/ stop word)	22	0	9	0	31	0
Delimitação Sintática	9/136	0	7/189	0	16	0
Erro de Segmentação	3/136	0	1/189	0	4	0

CAPÍTULO 5

DISCUSSÃO

Neste capítulo, discutem-se os resultados das anomalias na criação e manutenção do tesauro que foram apresentados no Capítulo 4.

Inicia-se o debate analisando as características da lista de discussão criada pelos lexicógrafos. Em seguida, discute-se cada tipo de anomalia: relacionamento, delimitação, permanência e tipo.

O processo de análise de anomalias teria uma melhor qualidade se fossem registradas passo a passo todas as alterações dos lexicógrafos, pois são perdidos alguns dados entre uma data da base de dados e outra, além de perder as transações do dia, dados esses que podem ser feitos e refeitos todos os dias, várias vezes pelos lexicógrafos.

Por se tratar de um trabalho inovador, não se encontrou na bibliografia uma pesquisa que possa servir de comparação.

5.1. LISTA DE DISCUSSÃO

Como fonte de avaliação desse estudo foi utilizada a lista de discussão criada a partir de *corpora* comparáveis, como descrito em Resultados. Esta técnica detectou alguns tipos de erros (ANDRADE, 2006).

A lingüística de *corpus* representa nada mais que a vida real, a forma usual cotidiana da língua (MCENERY e WILSON, 1996). Por outro lado, a análise do contexto e estudo quantitativo dos fatos lingüísticos permite descrever com objetividade a variação desses fenômenos evidenciando os problemas sintáticos e semânticos (HABER, 1997).

5.2. ANOMALIA DE RELACIONAMENTO

Soergel (1997) afirmou que há limitação nos tesauros devido à falta de precisão semântica, criando ambigüidades na interpretação e, quando o resultado dessa estrutura é irregular, não pode ser previsto com uma análise anterior. Existindo uma semântica ambígua que os caracterize, torna-os não adequados para o seu fim. Nos Resultados, o problema de ambigüidade foi evidenciado pela freqüência da anomalia de relacionamento. Foi esse tipo de anomalia que apresentou o maior número de ocorrências dentre as avaliadas neste trabalho, com 146 casos registrados, sendo que 41,09% deles foram debatidos na lista de discussão. Esse foi também o único caso em que ocorreram repetições de anomalias, nas quais a relação entre o mesmo conjunto de classes foi várias vezes alterada, resultando em 93 anomalias para um conjunto de 23 classes.

Observando este dado, pode-se verificar que o inglês/português produziu 7,89% de anomalias debatidas, enquanto no inglês/alemão foram 14,47%, e as anomalias debatidas repetidas atingiram 3,95% no inglês/português, contra 10,53% no inglês/alemão.

Na tabela 9 dos Resultados, pode-se observar que houve 43 anomalias repetidas, sendo que as ocorrências de anomalias repetidas puderam ser observadas com maior incidência nas não debatidas, representando 13 conjuntos de classes. Já o inglês/alemão teve 7 e 3 do inglês/português.

Quanto aos problemas encontrados na lista de discussão, realizada pelos lexicógrafos, comparando com esta anomalia, foram encontrados 4 problemas em cada língua, sendo que 2 problemas “sem relacionamento” e “Dois conceitos (diferentes na mesma classe)” foram encontrados em ambas.

Para o problema de “sem relacionamento” no inglês/português, das 26 ocorrências na lista de discussão, foram encontradas 14 anomalias, sendo que 5 conjuntos de classes foram alterados, representando 19,23% desse problema; já em inglês/alemão, foram registradas 60 ocorrências e destas foram observadas 28 anomalias, sendo 13 conjuntos de classes que foram alteradas, representando 21,67% deste problema.

O problema identificado como “dois conceitos (diferentes na mesma

classe)” na relação das linguas inglês/português houve 3 ocorrências, sendo encontrada 1 dessas anomalias, representando 33%, já em inglês/alemão foram registradas 8 ocorrências, sendo também encontrada 1 anomalia, correspondendo a 12,5% deste tipo de problema.

Neste contexto, o problema de ambigüidade lexical de certa forma é freqüente pelo fato que em tesauro multilíngüe os léxicos tendem a ser ambíguos, pois as classes de equivalências são constituídas por léxicos que, por sua vez, estão nos mais diversos idiomas. Acontece que o sentido de um léxico em determinando idioma não representa completamente o sentido em outro idioma. Em alguns casos, há a necessidade de separá-los em grupos distintos de classes de equivalência e realizar a relação de “*has_sense*” devido ao conflito entre os sentidos dos léxicos em outro idioma.

5.3. ANOMALIA DE DELIMITAÇÃO

No que diz respeito à delimitação de *string*, nessa pesquisa não se encontrou nenhuma anomalia. A quantidade de alterações foi, de certa forma, pequena, tendo ocorrido 135 alterações. Como o sistema *MorphoSaurus* já possuía uma boa cobertura de lexemas, é importante afirmar que esta anomalia não foi caracterizada em um tesauro com uma grande quantidade de termos como o *MorphoSaurus*, mas como não houve esta anomalia, não se pode afirmar; entretanto, pode-se considerar como uma extensão da relevância lexical. Erros de segmentação decorrentes de problemas de delimitação normalmente são remediados por meio da inclusão de novas variantes (e.g. *nephr* + *onephr* + *nephro*), o que, em si, não constitui anomalia. Também os lexicógrafos foram treinados no sentido de proceder de forma conservadora ao encontrar variantes de lexemas, mesmo exibindo padrões de delimitação pouco comuns.

5.4. ANOMALIA DE PERMANÊNCIA

É de suma importância que o tesauro possua somente lexemas relevantes. Como característica especial dos lexemas tipo *subword*, a delimitação de cada *string* que constitui um lexema desempenha um papel importante, pois trata-se basicamente de usar descritores representando um conjunto de lexemas para anotar documentos. Em virtude do desempenho de métodos de IR, a interferência por entidades lexicais que não possuem representatividade, efetividade e relevância no momento da recuperação do documento deve ser evitada.

Existem várias maneiras de se selecionar as palavras que representam os documentos, tais como *stopwords*, que são palavras muito freqüentes, porém de baixo significado, sendo inúteis para representar e distinguir documentos (FOX, 1992; KORFHAGE, 1997; RIJSBERGEN, 1979). Considera-se como *stopwords*: preposições, artigos, conjunções e demais palavras utilizadas para auxiliar na construção sintática das orações. Além dessas, também podem ser consideradas palavras específicas do contexto da coleção de documentos em questão (FOX, 1992).

Estudos que analisam a freqüência desses termos nos documentos observando o ponto de vista de RI, com a idéia principal de possuir palavras adequadas para criação de índices que facilitem a recuperação de documentos pertinentes. Esse tipo de pesquisa ainda não é uma questão definitiva, pois existem estudos que indicam que esses tipos de palavras não são úteis (NG, 1997; SALTON e BUCKLEY, 1987; SALTON e MACGILL, 1983). Contudo, existem estudos que indicam que tais palavras são relevantes (RILOFF, 1995).

Nesta pesquisa, o resultado de permanência mostra que das 5 anomalias avaliadas, 4 foram discutidas na lista de discussão e consideradas como falta de lexema ou classe; nesse caso, houve detecção de que esses lexemas são importantes para a recuperação de documentos, pois a não existência de um lexema no tesauro sendo que exista alguma ocorrência do mesmo no documento, este documento não será recuperado. Por outro lado, se este não possuir nenhuma ocorrência entre os documentos ele estará somente ocupando espaço e tempo de processamento. Contudo, a importância de se ter

esses lexemas no tesauro permite que o tesauro aumente sua abrangência.

Estudos mostram que a inclusão de novos lexemas é de extrema importância para aumentar a abrangência na RI, mas adverte que a inclusão de lexemas não pertinentes aumenta o número de documentos não relevantes recuperados. É pior para um SRI perder um bom documento do que ganhar novos documentos (VOORHEES, 1998).

Nesse caso, constatou-se na pesquisa que esta anomalia ocorreu somente nas relações entre as línguas inglês/português e inglês/alemão; já nas inglês/português e inglês/alemão, não houve ocorrência.

5.5. ANOMALIA DE TIPO

Considerando a anomalia de tipo, existem algumas técnicas que são debatidas, como a modelagem por conceitos, em que se propõe a análise de documentos, com a finalidade de fazer um mapeamento das palavras neles contidas, de forma a criar um “espaço conceitual” (LOH, 2001). Nesse espaço são utilizados métodos estáticos para determinar a delimitação conceitual das palavras. No método mais simples, realiza-se a análise de freqüência de todas as palavras contidas em um *corpus* se emprega. O processo é realizado automaticamente sem intervenção humana, onde aplica-se um método de corte, em que todas as palavras, abaixo dos limiares mínimos e máximos de freqüência, são descartadas (SALTON e MACGILL, 1983). Neste caso, podem ser perdidas algumas palavras importantes.

Outra técnica apresenta um algoritmo automático sem interação humana de reconhecimento e aquisição de sinônimos através de análise de freqüência de palavras, em resultados de consultas, é realizada pelo *Altavista* (http://www.altavista.com) (TURNEY, 2001).

O maior problema com essas estruturas construídas automaticamente é que elas utilizam técnicas baseadas na freqüência da ocorrência das palavras nos documentos e acabam não sendo muito eficientes (TURNEY, 2001).

Considerando a quantidade de anomalias encontradas, o inglês/português apresentou 7 anomalias, equivalentes a 38,89% deste tipo de anomalia, contra

11 anomalias do inglês/alemão, que representa 61,11%. Comparando com os problemas da lista de discussão, encontrou-se este tipo de anomalia em dois tipos de problemas “mesmo conceito em duas classes diferentes” e “sem relacionamentos”.

Nesta comparação, os valores ficam mais próximos, os problemas identificados como “mesmo conceito em duas classes diferentes” representam 11,11% no inglês/português, enquanto no inglês/alemão apresentou 11,63%, e os problemas “sem relacionamento” no inglês/português formam 15,38% e no inglês/alemão, 10%.

5.6. TRABALHOS FUTUROS

O conhecimento produzido a partir deste trabalho permitiu evocar novos aspectos, dos quais citam-se como possíveis trabalhos futuros:

(1) criação de um registro de procedimentos a cada evento, onde não se perderiam as alterações executadas no banco de dados que, neste trabalho, foi perdido;

(2) a partir da análise dos resultados desta dissertação, gerar mensagens em tempo real, que ajudariam a ter um melhor refinamento, pois apontam o problema, dando apoio aos lexicógrafos, agilizando ainda a manutenção do tesauro, pois, desse modo, os lexicógrafos podem discutir sobre o que há de errado no momento em que a mensagem aparece, auxiliando para que não haja retrabalho, como verificado na tabela 8, em que 10 classes denotaram mais de uma alteração desnecessária e onde poderia ter sido evocada a mensagem na primeira vez que foi emitida;

(3) criação de uma ferramenta de auditoria, cuja idéia principal seria bloquear alterações no tesauro. Um administrador do tesauro poderia ver esse sistema de mensagem e se não houvesse mais a necessidade de alteração, bloquearia de forma a não permitir alterações errôneas;

(4) criação de relatórios para verificação dos procedimentos existentes dentro do tesauro, pois a partir das tabelas implementadas para analisar as anomalias, são registrados todos os procedimentos.

CAPÍTULO 6

CONCLUSÕES

Devido aos fenômenos lingüísticos e aos aspectos semânticos e lexicográficos inerentes ao processo de representação do conhecimento no tesauro, há um dinamismo na manutenção de um tesauro envolvendo correções, muito freqüentes no início da construção do tesauro, ou redefinição de relacionamentos, quando o tesauro já possui uma quantidade suficiente de descritores que seja representativa da linguagem do domínio em questão.

Neste caso, constatou-se que o registro de procedimentos auxilia no refinamento do léxico, pois o *MorphoSaurus* já possui uma quantidade suficientemente representativa de descritores.

6.1. ANOMALIA DE RELACIONAMENTO

Verificou-se tanto a ocorrência de anomalia de relacionamento como também o processo de instabilidade de relacionamentos ocorrido no tesauro, tanto de especialização (desfazendo ambigüidades de lexemas ambíguos) quanto de composição (decompondo lexemas compostos). Conclui-se que além de identificar anomalias devido à instabilidade semântica, a metodologia desenvolvida ainda encontrou problemas de repetição da mesma anomalia, ocorrendo até 7 vezes a mesma anomalia.

6.2. ANOMALIA DE PERMANÊNCIA

Analisando os procedimentos de inclusão e exclusão de lexemas,

identificou-se a anomalia de permanência. Conclui-se que esse tipo de anomalia é reduzido, pelo fato de se tratar de um tesauro consolidado como é o caso do sistema *MorphoSaurus*; porém, em anomalia aplicada em um tesauro em nível inicial, a freqüência de transações de inclusão e exclusão será maior, possibilitando o aumento do índice desta anomalia, consequentemente.

6.3. ANOMALIA DE DELIMITAÇÃO

A anomalia de delimitação não ocorreu nenhuma vez nessa pesquisa, sendo a única que não foi detectada. Contudo, é uma anomalia existente em tesauros, e nessa pesquisa a quantidade de edições de lexemas foi somente de 135 modificações.

6.4. ANOMALIA DE TIPO

Observando as equivalências ocorridas durante a pesquisa, a identificação de anomalia de permanência foi a única categoria em que todos os casos anômalos ocorreram na lista de discussão.

6.5. CONSIDERAÇÕES FINAIS

Trabalhar com LN em sistemas computacionais é uma tarefa complexa, haja vista que fenômenos lingüísticos são responsáveis principalmente por interpretações ambíguas de ordem sintática e semântica. Apesar de haver um forte crescimento na área de PLN, este trabalho mostrou, pelos resultados obtidos, que muito ainda tem por se resolver quando se trata de construir um

tesauro de forma manual e, mais ainda, de forma automatizada.

O presente trabalho mostrou-se importante na construção de um tesauro de forma a servir de instrumento no auxilio aos lexicógrafos, pelo fato que, além de identificar as anomalias, esse método classifica qual o tipo de anomalia ocorrida.

Apesar de se ter trabalhado com um tesauro considerado consolidado aos olhos da técnica de validação e cobertura (*precision versus recall*), ainda assim foi possível, através desta abordagem, verificar anomalias de procedimentos realizados pelos lexicógrafos para verificação de uma amostra limitada de descritores semânticos.

Entre os tipos de anomalias, o que mais chamou atenção refere-se à instabilidade de relações semânticas, principalmente no que diz respeito à relação entre (conjuntos de) entidades lexicais ambíguas e os seus significados.

Esse fato mostra a verdadeira necessidade de expor dados aos lexicógrafos que apontam para repetições de procedimentos evitando, desta forma, falta de produtividade na construção e manutenção de um tesauro e, conseqüentemente, na qualidade do desempenho de um SRI como um todo.

A abordagem proposta mostrou-se viável para a monitoração da manutenção, mas acredita-se que uma monitoração contínua, ou seja, o registro de todos os procedimentos possa melhorar ainda mais os dados apresentados, pois o experimento foi realizado com bases de dados "discretas" (*backups*), podendo ser realizado com um monitoramento constaste.

A proposta mostra-se útil para a construção e monitoração de qualquer tesauro uma vez que fenômenos lingüísticos são inerentes ao processo. Pelo fato de se tratar com questões subjetivas como resolver ambigüidades, que podem levar aos erros de procedimentos e outros aspectos oriundos de fenômenos lingüísticos, essa metodologia constitui-se numa promissora ferramenta para amenizar o processo de gerenciamento do tesauro no que diz respeito à sua monitoração.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, R. L. Detecção de Erros em Tesauros Médico Multilíngüem através de Corpora Comparáveis. **Dissertação de Mestrado (Recuperação de Informação Multilíngüe em Saúde),** 113 f. CPGEI – Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, UTFPR, Curitiba, 2006.

APTÉ, C., DAMERAU F., WEISS S. M.. Automated learning of decision rules for text categorization. **ACM Transactions of Information Systems**, v.12, p.233-251, 1994.

ARAMPATZIS, A., WEIDE T., KOSTER C. Linguistically-motivated Information Retrieval. **Encyclopedia of Library and Information Science**, v.69, p.201-222, 2000.

BAEZA-YATES, R., RIBEIRO-NETO, B. **Modern Information Retrieval.** 1th Edition. New York: Addison Wesley Longman Publishing Co, 1999.

BEARDON, C. L., HOLMES, G. **Natural Language and Computational Linguistics.** Melksham-Wiltshire, England: Ellis Horwood Ltd, 1991.

BELKIN, N.J., CROFT W. B., Information Filtering and Information Retrieval: Two Sides of the Same Coin?, **ACM Transactions of Information Systems,** 35(12): 29-38, 1992.

BIDERMAN, M. T. Conceito Lingüístico de Palavra. In Basílio, M. (org.) **Palavra n°5.** Rio de Janeiro, Departamento de Letras da PUC, p. 81-97, 1999.

CARVALHO, E. C. A natureza social da Ciência da Informação. L. V. R. Pinheiro (Eds). **Ciência da Informação, Ciências Sociais e interdisciplinaridade.** Rio de Janeiro: IBICT, p. 51-63, 1999.

CHANKRAVARTHY, A. S., HAASE, K. B. NetSerf: using semantic knowledge to find Internet information archives. Proceedings. **Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR'95)**, 1995.

CINTRA, A. M. M.; KOBASHI, N.Y.; LARA, M.L.G. de & TÁLAMO, M.F.G. **Para entender as linguagens documentárias.** 2ª Ed. São Paulo: Polis, 2002.

CROFT, W. B. **Advances in Information Retrieval.** London: Kluwer Academic Publishers, 2000.

CURRAS, E. **Tesauros: linguagens terminológicas.** 286 p. Brasília : IBICT, 1995.

DAHLBERG, I. Ontical Structures and Universal Classification. Bangalore: Sarada Ranganathan Endowment, 1978.

FIGUEIREDO, L. M. D. O conceito de relevância e suas implicações. **Ciência da Informação,** Rio de Janeiro, v.6, n.2, p.75-78, 1977.

FOSKETT, D. J. Thesaurus. **Reading in Information Retrieval,** p.111-134. Morgan Kaufmann, 1997..

FOX, C. Lexical analysis and stoplists. In: FRAKES, W. B.; BAEZA-YATES, R. A. (Ed.). **Information Retrieval: Data Structures & Algorithms.** Upper Saddle River, New Jersey: Prentice Hall PTR, p. 102-130, 1992.

FRANCONI, E. Description Logics for Natural Language Processing. Baader, F.; McGuinness, D. L.; Nardi, D.; Patel-Schneider, P. P. (editores) **Description Logics Handbook.** Cambridge: Cambridge University Press, Cap.18, 2001.

FREE DICTIONARY, T. The Free Dictionary. EUA. 2005. Disponível em: http://encyclopedia.thefreedictionary.com. Acessado em 15 de novembro 2005.

FREGE, G. Sobre o Sentido e a Referência. **Lógica e Filosofia da Linguagem**. São Paulo, 1982.

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

FUHR N. and BUCKLEY C., Probabilistic Document Indexing from Relevance Feedback Data, **Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, p. 45-61, 1990.

FURNAS, G. W. The vocabulary problem in human-system communication. **ACM Transactions of Information Systems**, v.11, n.30, November 1987.

GEMET. General Environmental Multilingual Thesaurus, 2005. Disponível em: http://www.eionet.eu.int/gemet. Acesso em 26 setembro de 2005.

GOMES, H. E. **Manual de Elaboração de Tesauros Monolíngües.** Brasília, Programa Nacional de Bibliotecas das Instituições de Ensino Superior, 1990.

GONZALEZ, Marco A. I. Thesauri. (Trabalho Individual III, Pós-Graduação em Ciência da Computação, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul), 2001.

HARMAN, D. How effective is suffixing? **Journal of the American Society for Information Science.** v.1, n.42, p.7-15, 1991.

HABER, B. **Les Linguistiques de Corpus**, p.183. Paris : Armand Colin, 1997.

HERSH W. R. **Information Retrieval – A Health Care Perspective.** Computers and Medicine. New York: Springer, 1996.

JACQUEMIN, C. K., TZOUKERMANN, E., Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. **35 th Annual Meeting of the Association for Computational Linguistic (ACL) and 8th Conference of the European Chapter of the ACL**, Madri, p.24-31, 1997.

JACQUEMIN, C.; TZOUKERMANN, E. NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. Strzalkowski, Tomek (Ed.). **Natural Language Information Retrieval.** Kluwer Academic Publishers, p.25-74, 1999.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing – An Introduction to Natural Language Processing,** Computational Linguistics, and Speech Recognition. New Jersey, USA: Prentice-Hall, 934 p, 2000.

KENT, A., BERRY, M., LEUHRS, F.U., and PERRY, J.W. Machine literature learning VIII. **Operational criteria for designing information retrieval systems.** American Documentation, USA, v.6, n.2, p.93-101, 1955.

KORFHAGE, R. R. **Information Retrieval and Storage.** New York: John Wiley & Sons, 349 p, 1997.

KOWALSKI, G. **Information Retrieval Systems: Theory and Implementation.** Kluwer Academic Publishers, 282p, 1997.

KRIEGER, Maria da Graça. Terminologia técnico-científica: políticas lingüísticas e Mercosul. **Ciência e Cultura**, v.58, n.2, p.45-48, 2006.

KROVETZ, R. e CROFT, B. W. Lexical ambiguity and Information Retrieval. **ACM transaction on Information System,** v.10, n. 2, p.115-141, 1992.

KROVETZ, R. Homonymy and Polysemy in Information Retrieval. **Proceedings of the 35th Annual Meeting of the Association for Computacional Linguistics**, p.72-79, 1997.

KUSHMERICK, N. e THOMAS, B. Adaptive information extraction: Core technologies for information agents, **Lecture Notes in Computer Science**, v. 2586, p. 79-103, 2003.

LAFFLING, J. **On Constructing a transfer dictionary for man and machine**, v.4, n.1, p.17-31. Target, New York, 1992.

LANCASTER, F. W. **Indexação e resumos: teoria e prática.** Brasília : Briquet de Lemos/Livros, 1993.

LARA, Marilda Lopez Ginez de. Conceptual differences on terms and definitions and implications to the documentary language. **Ciência da Informação.** Vol.33, no.2, p.91-96, 2004.

LEWIS, David D.; SPARCK-JONES, Karen. Natural language processinf for information retrieval. **ACM Transactions of Information Systems,** v.39, n.1, 1996.

LOH, S. Abordagem baseada em conceitos para descoberta de conhecimento em textos. **Tese Doutorado em Ciência da Computação.** 110 f. Instituto de Informática, UFRGS, Porto Alegre, 2001.

MARKÓ K, DAUMKE P, SCHULZ S, HAHN U. Cross-Language MeSH indexing using morphsemantic normalization. In: **Proc AMIA Symposium**. 2003.

MCENERY, T. e WILSON, A. **Corpus Linguistics.** Edimburgo: Edinburgh University Press, 1996.

MEADOW, C. T.; Boyce, B. R.; Kraft, d. H. **Text Information Retrieval Systems.** 2th Edition. San Diego: Academic Press, 364 p, 2000.

MILLER, U. Thesaurus construction: problems and their roots. **Information Processing & Management,** v.33, n.4, p.481_493, Julho, 1997.

MILSTEAD, J. L. Use of thesauri in the full-text environment. **Indian Head, MD**, The Jelem Company, 1998. Disponível em: <http://www.bayside-indexing.com/Milstead/useof.htm> Acesso em 15 Novembro de 2005.

MOTTA, D. F. Método Relacional como Nova Abordagem para a Construção de Tesauros. **SENAI/DN/DPEA**, Rio de Janeiro, 1987.

NG, H. T. Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization. **Annual International ACM-SIGIR Conference On Research And Development In Information Retrieval**. Proceedings... New York: ACM Press, p.67-73, 1997.

NIE, J. Y., M. SIMARD, P. I, Durand R.. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. **Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval**, Berkeley, CA, USA, p. 74–75, 1999.

NLM - National Library of Medicine, Bethesda, MA, USA. **MeSH - Tree Structures & Alphabetic List**, 12th edition, Janeiro 2000.

NOGUEIRA G. N. N., ANDRADE R. L., MARKÓ K., NOHAMA P., SCHULZ S.. Recuperação Translingual de Textos via Representação Interlingual. **Congresso Brasileiro de informática em Saúde**, p.1202-1207, 2004.

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.

OLIVEIRA, R. P. **Uma história de delimitações teóricas: trinta anos de semântica no Brasil.** São Paulo: DELTA, v. 15, 1999.

OLIVEIRA, D. H. **Introdução a XML e suas aplicações**, 2002. Disponível em:. http://www.xml.com.br/docs/intro_xml_apli.pdf. Acesso em: 20 de janeiro 2006.

ORENGO, V. M. HUYCK, C. R. A Stemming algorithm for the Portuguese Language. **Proceedings of SPIRE´2001 Symposium on String Processing and Information Retrieval**, Laguna de San Raphael, Chile, 2001.

RIJSBERGEN C. J. **Information Retrieval**. London: Butterworth, 1979. Disponível em: <www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em 25 de maio de 2005.

RILOFF, E. Little words can make big difference for text classification. **Annual International Acm-Sigir Conference On Research And Development In Information Retrieval, SIGIR**. New York: ACM Press, p.130-136, 1995.

SAINT-DIZIER P. On the Polemorphic Behavior of Word-senses. **Lingüística Computacional: Investigação Fundamental e Aplicações.** Lisboa: Edição Colibri, p.29-56, 1999.

SALTON, G.; MACGILL, M. J. **Introduction to Modern Information Retrieval**. New York: McGRAW-Hill, 448 p, 1983.

SALTON, G.; BUCKLEY, C. **Term weighting approaches in automatic text retrieval.** Ithaca, New York: Department of Computer Science, Cornell University, 1987.

SANTOS, D. **DISPARA, a system for distributing parallel corpora on the Web.** In Nuno Mamede & Elisabete Ranchhod (eds.), Portugal for Natural Language Processing (PorTAL 2002), p. 23-26, 2002.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, p .41-62, 1996.

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

SCHULZ S., Nohama P., Borsato E. P., Matias L. J. D. Indexação e Recuperação Automática de Textos Médicos. **Congresso Brasileiro de Informática em Saúde - CBIS**, Natal, RN, 2000.

SCHULZ S., HONECK M, HAHN U: Biomedical Text Retrieval in Languages with a Complex Morphology. **Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain**. Philadelphia: Association for Computational Linguistics, p. 61-68, 2002.

SCHULZ S., NOGUEIRA G. N. N., ANDRADE R. L., MARKÓ K., NOHAMA P. Recuperação Translingual de Textos via Representação Interlingual. **Anais do Congresso Brasileiro de Informática em Saúde - CBIS**, Ribeirão Preto, SP, 2004.

SILVEIRA, M. L. Recuperação vertical de informação: um estudo de caso na área jurídica. **Tese de Doutorado**. Universidade Federal de Minas Gerais, Belo Horizonte, 2003.

SILVERSTEIN C., Analysis of a Very Large Web Search Engine Query Log, **Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, v.33, n.1, p. 6-12, 1999.

SMEATON, A. Information Retrieval: Still Butting Heads with Natural Language Processing?, **Information Extraction.** M.T Pazienza (Ed.), Springer-Verlag Lecture Notes in Computer Science (in press), 1997.

SOERGEL, D. Functions of a thesaurus – classification, ontological knowledge base. **College of Library and Information Services. University of Maryland**, 1997. Disponível em: <http://www.clis.umd.edu/faculty/soergel/soergelfctclass.pdf > Acesso em 26 de Setembro de 2005.

SOERGEL, D. Multilingual Thesauri in Cross-language Text and Speech Retrieval. **AAAI Symposium on Cross-Language Text and Speech Retrieval**, p.1-8, 1997.

SPARCK-JONES, K.; WILLET, P. (editores). **Readings in information retrieval**. California: Morgan Kaufmann Publishers, Inc., 1997.

STREHL, L. Evaluation of indexing consistency in an arts university library. **Scielo**, v. 27, n. 3, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000300011&lng=en&nrm=iso>. Acesso em 06 de Agosto de 2006.

SUN, Microsystems. Java Programming Language, 1995. Disponível em http://java.sun.com. Acesso em 21 Setembro 2004.

SUN, Microsystems. JDBC API, 2001. Disponível em http://java.sun.com/products/jdbc. Acesso em 21 Setembro 2004.

SUN, Microsystems. JavaServer Pages. V.1.2. 2001a. Disponível em http://java.sun.com/products/jsp. Acesso em 21 Setembro 2004.

THE FREE DICTIONARY, 2005. Disponível em: http://encyclopedia.thefreedictionary.com/. Acesso em 15 de Novembro de 2005.

TURNEY, P. D. Mining the WEB for synonyms: PMI-IR versus LSA on TOEFL. **European Conference On Machine Learning (ECML). Lecture Notes in Computer Science, 2167**. Heidelberg, Germany, p.491-502. 2001.

UMLS. Knowledge Sources 5th Experimental Edition, **Unified Medical Language System - U.S. Departament of Health and Human Services, National Institutes of Health, National Library of Medicine**, 1994.

VERSTRAETE, T. Entrepreneuriat: modélisation du phénomène. **Revue del´Entrepreneuriat**, v.1, n.1, p. 20, 2001.

VOORHEES E. M.; HARMAN D. Overview of the fifth Text Retrieval Conference. **Proceedings of the fifth Text Retrieval Conference (TREC-5)**, p. 1-28, 1996.

VOORHEES, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. **Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98).** Melbourne, Australia: ACM PRESS, p.1-28, 1998.

WANG, P. Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. **Information Processing and Management**. 36, p. 229-251, 2000. Disponível em: <https://206.191.28.118/docushare/dsweb/Get/Document-1442/Wang_et_al_ 2001 _IR_model.pdf>. Acesso em Novembro de 2005.

WILLIE, S; BRUZA P. Users Model of the Information Space: the Case for Two Search Models. **Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '95)**. p 205 – 211, 1995.

XML: Extensible Markup Language (2001). Disponível em <http://www.w3.org/tr/rec-xml>. Acesso em: 21 de Setembro 2004.

YATES, R. B. String Searching Algorithms. FRAKES,William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992.

YATES, R. B. An extended model for full text databases. **Journal of the Brasilian Computer Society**, v.2, n.3, Abril 1996.

YATES, R. B.; RIBEIRO, B. A. **Modern Information Retrieval.** Addison Wesley, 1999.

ANEXO

AS 146 ANOMALIAS
REFERENTE A DELIMITAÇÃO SEMÂNTICA

NÚMERO DA ANOMALIA	EQ CLASS 1	EQ CLASS 2	AÇÃO	DATA	LINGUA	PROBLEMA RELACIONADO NA LISTA DE DISCUSSÃO
1	162	503699	sense_of	2005 06 14		sem problemas
2	237	2471	sense_of	2005 07 13		sem problemas
3	247	18819	sense_of	2005 06 30	português	Falta de lexema ou classe
4	344	500982	has_sense	2005 10 06		sem problemas
5	646	502731	sense_of	2005 10 04	alemão	sem relacionamento
6	1285	502538	has_sense	2005 06 30		sem problemas
7	1473	502563	sense_of	2005 09 16	alemão	sem relacionamento
8	1605	501428	word_part_of	2005 06 10		sem problemas
9	1642	502815	has_word_part	2005 06 30		sem problemas
10	1642	504079	has_word_part	2005 06 30		sem problemas
11	2217	501635	sense_of	2005 08 08	português	sem relacionamento
12	2431	501638	sense_of	2005 08 05		sem problemas
13	2431	504759	sense_of	2005 08 05	alemão	Dois conceitos (diferentes) na mesma classe
14	2958	3136	has_sense	2005 10 10		sem problemas
15	2958	500783	sense_of	2005 10 10	português	Problemas ortográficos
16	3757	36746	has_sense	2005 06 10		sem problemas
17	3830	25824	sense_of	2005 07 25		sem problemas
18	3913	119	word_part_of	2005 07 28		sem problemas
19	3913	1921	word_part_of	2005 07 28		sem problemas
20	4555	608	sense_of	2005 06 10	português	sem relacionamento
21	9697	24772	sense_of	2005 09 28	alemão	sem relacionamento
22	10020	505659	sense_of	2005 10 10		sem problemas
23	10567	500244	sense_of	2005 08 25		sem problemas
24	13525	1605	has_word_part	2005 06 10		sem problemas
25	14483	502707	sense_of	2005 08 18	alemão	sem relacionamento
26	14561	17468	has_sense	2005 07 25		sem problemas
27	14737	10364	has_word_part	2005 06 30	português	Falta de lexema ou classe
28	17896	502752	sense_of	2005 05 31		sem problemas
29	19226	12029	sense_of	2005 07 19		sem problemas
30	35972	502538	has_sense	2005 06 30		sem problemas
31	36079	10155	has_sense	2005 07 12		sem problemas
32	38207	94	has_sense	2005 09 28	alemão	sem relacionamento
33	500474	2583	has_sense	2005 06 30		sem problemas
34	500566	10154	has_word_part	2005 05 31		sem problemas
35	500912	3925	has_sense	2005 10 12		sem problemas
36	501020	512	has_sense	2005 07 26		sem problemas

37	501105	28190	has_sense	2005 05 31	português	Dois conceitos (diferentes) na mesma classe
38	501981	35972	sense_of	2005 06 30		sem problemas
39	502031	2239	has_sense	2005 09 26		sem problemas
40	502488	24772	has_sense	2005 09 28	alemão	sem relacionamento
41	503396	24772	sense_of	2005 09 28	alemão	sem relacionamento
42	503452	5819	has_sense	2005 06 08		sem problemas
43	503452	15092	has_sense	2005 06 08		sem problemas
44	503629	1605	has_sense	2005 06 10		sem problemas
45	503629	503624	has_sense	2005 06 10		sem problemas
46	504078	35972	sense_of	2005 06 30		sem problemas
47	504081	28771	has_sense	2005 06 30		sem problemas
48	504741	15212	has_sense	2005 08 04	alemão	sem relacionamento
49	505113	3476	has_sense	2005 08 26	alemão	mesmo conceito em duas classes
50	505659	1387	has_sense	2005 10 10		sem problemas
51	505661	10020	has_sense	2005 10 10		sem problemas
52	4178	501789	has_word_part	2005 08 22	alemão	sem relacionamento
53	9393	4073	has_sense	2005 05 31		sem problemas
54	2499	505317	sense_of	2005 08 08	alemão	Termo especifico do idioma
55	2499	505317	word_part_of	2005 09 28	alemão	Termo especifico do idioma
56	938	504023	has_sense	2005 06 22		sem problemas
57	938	504023	sense_of	2005 06 22		sem problemas
58	500421	504023	has_sense	2005 06 22		sem problemas
59	500421	504023	sense_of	2005 06 22		sem problemas
60	504050	501658	sense_of	2005 06 23		sem problemas
61	504050	501658	has_sense	2005 07 11		sem problemas
62	505317	2499	has_sense	2005 09 28	alemão	Termo especifico do idioma
63	505317	2499	has_word_part	2005 09 28	alemão	Termo especifico do idioma
64	2035	2217	has_sense	2005 08 04	alemão	sem relacionamento
65	2217	2035	sense_of	2005 08 08	alemão	sem relacionamento
66	505670	3	sense_of	2005 09 28	português	sem relacionamento
67	505670	3	sense_of	2005 10 10	português	sem relacionamento
68	3	505670	has_sense	2005 10 11	português	sem relacionamento
69	3	505670	has_sense	2005 10 12	português	sem relacionamento
70	4656	20047	has_word_part	2005 06 09		sem problemas
71	4656	20047	word_part_of	2005 06 09		sem problemas
72	20047	4656	word_part_of	2005 06 09		sem problemas
73	20047	4656	has_word_part	2005 06 09		sem problemas
74	574	10032	word_part_of	2005 06 13	alemão	mesmo conceito em duas classes
75	574	10032	word_part_of	2005 07 18	alemão	mesmo conceito em duas classes
76	10032	574	has_word_part	2005 08 08	alemão	mesmo conceito em duas classes
77	10032	574	has_word_part	2005 08 22	alemão	mesmo conceito em duas classes
78	1820	10032	has_word_part	2005 08 22		sem problemas
79	1820	10032	word_part_of	2005 08 22		sem problemas
80	10032	1820	word_part_of	2005 08 22		sem problemas
81	10032	1820	has_word_part	2005 08 22		sem problemas
82	800	12012	has_sense	2005 10 04	português	sem relacionamento
83	800	12012	has_sense	2005 10 10	português	sem relacionamento
84	12012	800	sense_of	2005 10 11	português	sem relacionamento
85	12012	800	sense_of	2005 10 12	português	sem relacionamento

86	2104	25222	has_sense	2005 07 18		Sem problemas
87	2104	25222	sense_of	2005 07 18		sem problemas
88	25222	2104	sense_of	2005 07 28		sem problemas
89	25222	2104	has_sense	2005 07 28		sem problemas
90	504619	501029	sense_of	2005 06 13		sem problemas
91	504619	501029	sense_of	2005 06 22		sem problemas
92	501029	504619	sense_of	2005 07 28		sem problemas
93	504619	501029	sense_of	2005 07 28		sem problemas
94	224	503250	word_part_of	2005 05 28		sem problemas
95	503250	224	has_word_part	2005 06 13		sem problemas
96	503250	224	sense_of	2005 06 22		sem problemas
97	224	503250	has_sense	2005 06 23		sem problemas
98	162	503552	sense_of	2005 06 13		sem problemas
99	503552	162	has_sense	2005 06 22		sem problemas
100	162	503552	sense_of	2005 06 23		sem problemas
101	503552	162	has_sense	2005 07 28		sem problemas
102	503552	850	has_sense	2005 06 13		sem problemas
103	850	503552	has_sense	2005 06 13		sem problemas
104	503552	850	has_sense	2005 07 28		sem problemas
105	850	503552	has_sense	2005 07 28		sem problemas
106	261	503940	sense_of	2005 05 28	português	sem relacionamento
107	261	503940	word_part_of	2005 05 28	português	sem relacionamento
108	503940	261	has_sense	2005 06 22	português	sem relacionamento
109	503940	261	has_word_part	2005 06 22	português	sem relacionamento
110	401	504874	sense_of	2005 08 04	alemão	Termo especifico do idioma
111	401	504874	word_part_of	2005 08 04	alemão	Termo especifico do idioma
112	504874	401	has_sense	2005 08 08	alemão	Termo especifico do idioma
113	504874	401	has_word_part	2005 08 08	alemão	Termo especifico do idioma
114	505279	505278	has_sense	2005 07 22		sem problemas
115	505279	505278	sense_of	2005 07 22		sem problemas
116	505278	505279	has_sense	2005 09 26		sem problemas
117	505278	505279	sense_of	2005 09 26		sem problemas
118	1837	655	sense_of	2005 08 08	alemão	sem relacionamento
119	1837	655	sense_of	2005 08 11	alemão	sem relacionamento
120	1837	655	sense_of	2005 09 26	alemão	sem relacionamento
121	1837	655	has_sense	2005 10 04	alemão	sem relacionamento
122	655	1837	has_sense	2005 10 04	alemão	sem relacionamento
123	201836	504945	sense_of	2005 08 04	alemão	Sem relacionamento
124	201836	504945	sense_of	2005 08 04	alemão	sem relacionamento
125	504945	201836	has_sense	2005 08 08	alemão	sem relacionamento
126	504945	201836	has_sense	2005 08 11	alemão	sem relacionamento
127	504945	201836	has_sense	2005 08 11	alemão	sem relacionamento
128	501029	501817	has_sense	2005 06 22		sem problemas
129	501029	501817	has_sense	2005 07 22		sem problemas
130	501029	501817	has_sense	2005 07 28		sem problemas
131	501817	501029	has_sense	2005 07 28		sem problemas
132	501817	501029	sense_of	2005 08 08		sem problemas
133	501817	501029	sense_of	2005 08 11		sem problemas
134	504536	24740	word_part_of	2005 07 22		sem problemas
135	504536	24740	has_word_part	2005 07 22		sem problemas
136	24740	504536	word_part_of	2005 07 22		sem problemas
137	24740	504536	has_word_part	2005 07 28		sem problemas
138	24740	504536	word_part_of	2005 08 08		sem problemas

139	24740	504536	word_part_of	2005 09 30		sem problemas
140	2217	504736	sense_of	2005 07 22	alemão	sem relacionamento
141	504738	2217	has_sense	2005 07 28	alemão	sem relacionamento
142	2217	504738	sense_of	2005 08 04	alemão	sem relacionamento
143	2217	504738	has_sense	2005 08 08	alemão	sem relacionamento
144	2217	504738	sense_of	2005 08 08	alemão	sem relacionamento
145	504738	2217	sense_of	2005 08 08	alemão	sem relacionamento
146	504738	2217	has_sense	2005 08 08	alemão	sem relacionamento

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