EDSON JOSÉ PACHECO

MorphoMap:

Mapeamento automático de narrativas clínicas para uma terminologia médica

Tese de doutorado apresentada ao programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de "Doutor em Ciências" – Área de Concentração: Engenharia Biomédica.

Orientador:
Prof. Dr. Percy Nohama

Co-Orientador:
Prof. Dr. Stefan Schulz

Curitiba,
Dezembro de 2009.

Dados Internacionais de Catalogação na Publicação

P116 Pacheco, Edson José

MorphoMap : mapeamento automático de narrativas clínicas para terminologia médica / Edson José Pacheco. — 2009.

154 f. : il. ; 30 cm

Orientador : Percy Nohama

Co-orientador : Stefan Schulz

Tese (Doutorado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Curitiba, 2009

Bibliografia : 137-153

1. Ontologia. 2. Sistemas de processamento da fala. 3. Linguística – Processamento de dados. 4. Sistemas de recuperação da informação. 5. Informática médica. 6. Medicina – Processamento de dados. 7. Registros médicos. 8. Engenharia elétrica – Teses. I. Nohama, Percy, orient. II. Schulz, Stefan, co-orient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial. IV. Título.

CDD (22. ed.) 621.3

Biblioteca Central da UTFPR, Campus Curitiba

Aos meus sábios pais, José e Adenilde, pelos exemplos de vida.

À minha amada esposa Marilena, com quem divido o prazer desta jornada chamada vida.

Ao meu filho Alexsandro, que cresça na certeza do meu amor incondicional, em um mundo renovado na busca da igualdade e fraternidade.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Campus Curitiba

Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial

Título da Tese Nº 47:

"MorphoMap: Mapeamento Automático de Narrativas Clínicas para uma Terminologia Médica"

por

Edson José Pacheco

Esta tese foi apresentada, às 08h30min do dia 16 de dezembro de 2009, como requisito parcial para a obtenção do título de DOUTOR EM CIÊNCIAS – Área de Concentração: Engenharia Biomédica, pelo Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial – CPGEI – da Universidade Tecnológica Federal do Paraná – UTFPR. O trabalho foi aprovado pela Banca Examinadora, composta pelos professores:

Prof. Dr. Percy Nohama
(Orientador - UTFPR)

Prof. Dr. Marco Antonio Gutierrez
(USP)

Prof. Dr. Flávio Bortolozzi
(CESUMAR)

Prof. Dr. Alcides Calsavara
(PUC-PR)

Prof. Dr. César Augusto Tacla
(UTFPR)

Visto da coordenação:

Prof. Dr. Humberto Remígio Gamba
(Coordenador do CPGEI)

Av. Sete de Setembro, 3165 - 80230-901 - Curitiba – PR.
Tel. +55 (41) 3310 4680 / e-mail: cpgei-ct@utfpr.edu.br / http://www.cpgei.ct.utfpr.edu.br

AGRADECIMENTOS

Nem sempre há oportunidade de agradecermos a todas as pessoas, as quais participaram em nossas vidas contribuindo para a nossa formação, seja de forma direta ou indireta. A escrita de uma tese apresenta esta oportunidade. Gostaria primeiramente de agradecer aos Professores Dr. Percy e Dr. Stefan, que mais do que supervisores ("*supervisor*") são verdadeiros orientadores de vida. Obrigado pela paciência, pelos ensinamentos e, acima de tudo, pelas oportunidades que, com certeza, em muitas ocasiões não foram merecidas. Sem o apoio deles eu não teria alcançado este sonhado objetivo.

O ambiente no qual se vive possui a característica de facilitar ou dificultar o alcance de objetivos. O LER (Laboratório de Engenharia de Reabilitação) e o *Institut für Medizinische Biometrie und Medizinische Informatikde* caracterizaram-se como ambientes estimulantes para a mente e alma. Quero registrar os meus agradecimentos ao Philipp, ao Kornel, ao Roosewelt, à Ana Carolina, à Priscila, ao Michel, ao Jeferson, ao Píndaro e à Cláudia, que estiveram envolvidos no projeto e na pesquisa. Obrigado a todos.

Agradeço às Professoras Dra. Mariza Kluck e Dra. Cláudia Moro, pelo especial e fundamental envolvimento nas etapas iniciais da pesquisa.

Agradeço ao CNPq e ao DRL, que financiaram parte da pesquisa.

Agradeço ao HCPA e à PUCPR, pelo apoio aos trabalhos realizados.

Complexas de serem mensuradas, porém mais fáceis de serem sentidas, assim são as relações familiares. Agradeço aos meus pais, que mesmo com formação formal incompleta, nunca deixaram de inspirar o desenvolvimento de seus filhos. Saibam que esta vitória é especialmente de vocês.

Por fim, gostaria de deixar registrado meu amor e apreço por minha esposa, pelo seu apoio incondicional, e ao meu ainda não nascido filho Alexsandro: saibas que a confirmação de sua existência trouxe ânimo e norte para seu pai. Obrigado!

Nestes cinco anos de trabalho muitos foram os que estiveram ao meu lado, participando desta jornada, certamente estes poucos parágrafos não foram suficientes para citar e agradecer a todos. Portanto, peço desculpas aos que aqui não estão presente, saibam, no entanto, que vocês fazem parte do meu pensamento e da minha eterna gratidão.

“O conhecimento deve ser permanentemente revisitado e revisado pelo pensamento.” (Morin, 2001)

RESUMO

PACHECO, Edson. MorphoMap: Mapeamento automático de narrativas clínicas para uma terminologia médica. 2009. 154 f. Tese – Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2009.

A documentação clínica requer a representação de situações complexas como pareceres clínicos, imagens e resultados de exames, planos de tratamento, dentre outras. Entre os profissionais da área de saúde, a linguagem natural é o meio principal de documentação. Neste tipo de linguagem, caracterizada por uma elevada flexibilidade sintática e léxica, é comum a prevalência de ambigüidades em sentenças e termos. O objetivo do presente trabalho consiste em mapear informações codificadas em narrativas clínicas para uma ontologia de domínio (SNOMED CT). Para sua consecução, aplicaram-se ferramentas processamento de linguagem natural (PLN), assim como adotaram-se heurísticas para o mapeamento de textos para ontologias. Para o desenvolvimento da pesquisa, uma amostra de sumários de alta foi obtida junto ao Hospital das Clínicas de Porto Alegre, RS, Brasil. Parte dos sumários foi manualmente anotada, com a aplicação da estratégia de *Active Learning*, visando a preparação de um *corpus* para o treinamento de ferramentas de PLN. Paralelamente, foram desenvolvidos algoritmos para o pré-processamento dos textos 'sujos' (com grande quantidade de erros, acrônimos, abreviações, etc). Com a identificação das frases nominais, resultado do processamento das ferramentas de PLN, diversas heurísticas (identificação de acrônimos, correção ortográfica, supressão de valores numéricos e distância conceitual) para o mapeamento para a SNOMED CT foram aplicadas. A versão atual da SNOMED CT não está disponível em português, demandando o uso de ferramentas para processamento multi-lingual. Para tanto, o pesquisa atual é parte da iniciativa do projeto MorphoSaurus, por meio do qual desenvolve-se e disponibiliza-se um *thesaurus* multi-língue (português, alemão, inglês, espanhol, sueco, francês), bem como componentes de software que permitem o processamento inter-lingual. Para realização da pesquisa, 80% da base de sumários foi analisada e manualmente anotada, resultando em um *corpus* de domínio (textos médicos e em português) que permitiu a especialização do software OpenNLP (baseado no modelo estatístico para o PLN e selecionado após a avaliação de outras soluções disponíveis). A precisão do etiquetador atingiu 93.67%. O *thesaurus* multi-língue do MorphoSaurus foi estendido, reestruturado e avaliado (automaticamente com a comparação por meio de textos comparáveis – 'traduções de um mesmo texto para diferentes idiomas') e sofreu intervenções objetivando a correção de imperfeições existentes, resultando na melhoria da cobertura lingüística, no caso do português, em 2%; e 50% para o caso do espanhol, medidas obtidas por meio do levantamento das curvas de precisão e revocação para a base do OHSUMED. Por fim, a codificação de informações de narrativas clínicas para uma ontologia de domínio é uma área de elevado interesse científico e clínico, visto que grande parte dos dados produzidos quando do atendimento médico é armazenado em texto livre e não em campos estruturados. Para o alcance deste fim, adotou-se a SNOMED CT. A viabilidade da metodologia de mapeamento foi demonstrada com a avaliação dos resultados do mapeamento automático contra um padrão ouro, manualmente desenvolvido, indicando precisão de 83,9%.

Palavras-Chave: Ontologia, SNOMED CT, *Information Retrieval*, Processamento de Linguagem natural, prontuário eletrônico.

ABSTRACT

PACHECO, Edson. MorphoMap: Automatic Mapping of Clinical Documentation to a medical terminology. 2009. 154 f. Tese – Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2009.

Clinical documentation requires the representation of fine-grained descriptions of patients' history, evolution, and treatment. These descriptions are materialized in findings reports, medical orders, as well as in evolution and discharge summaries. In most clinical environments natural language is the main carrier of documentation. Written clinical jargon is commonly characterized by idiosyncratic terminology, a high frequency of highly context-dependent ambiguous expressions (especially acronyms and abbreviations). Violations of spelling and grammar rules are common. The purpose of this work is to map free text from clinical narratives to a domain ontology (SNOMED CT). To this end, natural language processing (NLP) tools will be combined with a heuristic of semantic mapping. The study uses discharge summaries from the Hospital das Clínicas de Porto Alegre, RS, Brazil. Parts of these texts are used for creating a training *corpus*, using manual annotation supported by active learning technology, used for the training of NLP tools that are used for the identification of parts of speech, the cleansing of "dirty" text passages. Thus it was possible to obtain relatively well-formed and unambiguous noun phrases, heuristics was implemented to semantic mapping between these noun phrases (in Portuguese) and the terms describing the SNOMED CT concepts (English and Spanish) uses the technology of morphosemantic indexing, using a multilingual subword thesaurus, provided by the MorphoSaurus system, the resolution of acronyms, and the identification of named entities (e.g. numbers). In this study, 80 per cent of the summaries are analyzed and manually annotated, resulting in a domain *corpus* that supports the specialization of the OpenNLP system, mainly following the paradigm of statistical natural language processing (the accuracy of the tagger obtained was 93.67%). Simultaneously, several techniques have been used for validating and improving the subword thesaurus. To this end, the semantic representations of comparable test corpora from the medical domain in English, Spanish, and Portuguese were compared with regard to the relative frequency of semantic identifiers, improving the corpus coverage (2% to Portuguese, and 50% to Spanish). The result was used as an input by a team of lexicon curators, which continuously fix errors and fill gaps in the trilingual thesaurus underlying the MorphoSaurus system. The progress of this work could be objectified using OHSUMED, a standard medical information retrieval benchmark. The mapping of text-encoded clinical information to a domain ontology constitutes an area of high scientific and practical interest due to the need for the analysis of structured data, whereas the clinical information is routinely recorded in a largely unstructured way. In this work the ontology used was SNOMED CT. The evaluation of mapping methodology indicates accuracy of 83.9%.

Keywords: Ontologies, SNOMED CT, *Information Retrieval*, natural language processing, electronic medical records.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

CAS	*Common Analysis Structure*
CLIR	*Cross-Language Information Retrieval*
CFM	Conselho Federal de Medicina
CTV3	*Clinical Terms Version* 3
HCPA	Hospital das Clínicas de Porto Alegre
IA	Inteligência Artificial
IAD	Inteligência Artificial Distribuída
IDF	*Inverse Document Frequency*
IR	*Information Retrieval*
IV	Invariante
KIF	*Knowledge Interchange Format*
KMG	*Knowledge Modeling Group*
MID	*MorphoSaurus IDentifiers*
MS	Ministério da Saúde
NHS	*National Health Service*
NLP	*Natural Language Processing*
NP	*Nominal Phrase*
OOP	*Object-oriented programming*
PF	Prefixo
PLM	Processamento da Linguagem Médica
PLN	Processamento de Linguagem Natural
POO	Programação Orientada a Objetos
PP	Prefixo Próprio
PUCPR	Pontifícia Universidade Católica do Paraná
RDF	*Resource Description Framework*
SADT	Serviços Auxiliares de Diagnósticos e Terapia
SF	Sufixo
SFP	Sufixo Próprio
ST	*Stem*
SUS	Sistema Único de Saúde

TF	*Term Frequency*
UFRGS	Universidade Federal do Rio Grande do Sul
UIM	*Unstructured information management*
UIMA	*Unstructured Information Management Architecture*
UMLS	*Unified Medical Language System*
UTFPR	Universidade Tecnológica Federal do Paraná
UTI	Unidade de Terapia Intensiva
WSDL	*Web Services Description Language*
WWW	*World Wide Web*
XML	*EXtensible Markup Language*

SUMÁRIO

Capítulo 1

1 INTRODUÇÃO

1.1 Contexto

Os estudos antropológicos, tanto na linha evolucionária quanto do desenvolvimento filogenético, mostram que os homens têm uma tendência "espontânea" a "descobrir" o que é o mundo que os circunda, a conhecer, a buscar compreender o que é esse mundo (Konder, 2002; Minsky, 1986; Novak, 1977; Severino, 2002). Segundo Severino (2002), "conhecer é uma atividade que brota de um impulso original que se confunde, na sua gênese, com o próprio impulso da vida. Na verdade, a atividade consciente, subjetiva, emerge e se desenvolve integrando-se à atividade vital como um todo, o pensamento constituindo-se como processo imanente da ação humana visando a sua própria sobrevivência bio-material".

O conhecimento, neste contexto, configura-se como o esforço do "espírito" humano para compreender a realidade circundante. Esta compreensão ocorre mediante a interpretação dos sinais captados pelos sentidos. O significado, a partir da interpretação dos sinais, é resultante dos processos cognitivos que, por sua vez, acontecem mediante a associação entre símbolos representando objetos do mundo e a realidade observada. O elemento central desta subjetividade instauradora de relações entre os diversos aspectos da "realidade" é exatamente a capacidade de "duplicar" elementos da experiência humana através do processo de simbolização (representação mental), permitindo que a realidade circundante seja tratada no plano simbólico(Kodratoff, 1990; Minsky, 1986; Severino, 2002).

Enquanto representação primeira, a linguagem é o meio mais utilizado para a representação e disseminação de conhecimento, elemento essencial no processo de

simbolização. No domínio da disseminação, o primeiro grande paradigma foi o da dialética[1], substituído por novas técnicas e metodologias, apoiadas na antiga, visto que a escrita nasce com a consciência humana, e nova forma, no aspecto da facilidade na replicação, de registrar o circundante, quando do advento da impressão e tipografia, possibilitado pela invenção de Johannes Gutenberg (Smeaton, 1997).

Com a definição das bases, técnicas e conceituais, de comunicação e intercâmbio de informações e conhecimento, os últimos séculos foram caracterizados pelo acúmulo desses elementos, permitindo a construção de bibliotecas, de métodos de indexação para a localização da informação desejada, em síntese, com o crescimento da disponibilidade surge a necessidade da guarda e recuperação da informação (Marques, 2002; Shmeil, 1999; Wooldridge, 1995).

Com a adoção das tecnologias de informação, tais mecanismos, os de guarda e recuperação, sofreram profunda alteração nos últimos anos, propiciando a produção e manipulação de grandes volumes de dados, informações e conhecimento, virtualmente disponíveis a todos pelo uso das tecnologias de acesso, demandando novas técnicas de indexação, pois este cenário torna a tarefa de recuperação de informação difícil, custosa e muitas vezes complexa para os usuários (Baeza-Yates e Ribeiro-Neto, 1999; Chu e Rosenthal, 1996; Smeaton, 1997).

Neste sentido, o homem, durante o seu período de vida, está constantemente modificando o estado do mundo real. O registro dessas modificações é realizado com propósitos variados, no domínio das ciências médicas, as evidências (Minchin e Vangenot, 2006; Salem e Alfonse, 2008; Nytro e Sorby, 2009) apontam que os registros clínicos, quando analisados individualmente, em momento temporal isolado, sob a luz do conhecimento científico da época, pode induzir a conclusões parciais, dissociadas da realidade "efetivamente" presente. No mesmo domínio, a análise do quadro clínico de um paciente é multidisciplinar e multiaxial, um evento isolado, aparentemente sem valor, pode se tornar preponderante quando do inter-relacionamento com outras observações, realizadas por diferentes profissionais, em diferentes momentos, sob condições clínicas específicas.

Registrar, portanto, é condição *sine qua non* para a análise integral do quadro clínico de um paciente. Em tese, todo evento, independente da relevância "atual", pode auxiliar na tomada de decisão futura, quando de uma ocorrência médica (Kanou, Jouber e Maury, 1998). Qual é, portanto, o critério que define as boas práticas para a seleção dos eventos que merecem registro? A resposta passa pela definição dos processos de aprendizagem, que

[1] Alguns historiadores consideram Sócrates como fundador da dialética, já Aristóteles considerava Zênon de Eléa (Konder, 2002).

conduzem à identificação dos elementos necessários para as pesquisas e tomadas de decisão, em concomitância com as características técnicas da tecnologia adotada (Sivagurunathan e Chountas, 2003).

Identificados e enunciados os processos que compõem as boas práticas relacionadas ao registro, deve-se considerar que cada vez mais é reconhecido o fato de que a complexidade das áreas de assistência à saúde e ciências biológicas necessita de um consenso a respeito dos termos e linguagem utilizados em documentos e na comunicação. Tal necessidade é impulsionada pelo crescimento exponencial de dados gerados nos contextos da assistência ao paciente e das pesquisas biológicas que, segundo Pestonik (2000), crescem a uma taxa anual de 7%, e "a base de dados tende a dobrar, pelo menos, a cada 10 anos".

Mesmo quando disponível, os dados clínicos não "podem ser completamente explorados em termos de integração, recuperação ou interoperabilidade, porque os sistemas básicos de terminologia e classificação (freqüentemente classificados sob o tópico "terminologia biomédica" – conforme descrito na Tabela 1) são inadequados de diversas formas" (Schulz, Stenzhorn, *et al.*, 2009). Sua heterogeneidade reflete os diferentes propósitos e comunidades – incluindo aquelas à parte da tecnologia da informação – e cria um grave obstáculo à interoperabilidade e agregação consistentes de dados, conforme exigido pela pesquisa biomédica e a assistência à saúde, assim como para o uso destas informações para o apoio a toma de decisão (Lewis, 2002; Tretiakov, Hunter, *et al.*, 2006).

Tabela 1 — Instâncias de terminologias clínicas

Terminologia	**Propósito**
CID-9/CID-10 (WHO, 2009)	Classificação de doenças, estatísticas de saúde, faturamento hospitalar
Dicionário de Medicamentos da OMS (UMC, 2009)	Classificação de medicamentos
ATC (WHOCC, 2009)	
RxNorm (NLM, 2009)	
BRASINDICE (BRASINDICE, 2009)	
LOINC (LOINC, 2009)	Comunicação inter-laboratorial
DICOM (MITA, 2009)	Descrições de imagens médicas e processos de diagnóstico por *imagem*
MeSH(Nytro e Sorby, 2009)	Indexação da literatura médica
SNOMED CT (IHTSDO, 2009)	Documentação e codificação clínica
AMB (ABRALAPAC, 2009)	Classificação de exames e procedimentos
CBHPM (ABRALAPAC, 2009)	
TUEP – SUS (SUS, 2009)	
TUSS (CNS, 2009)	

Como solução intermediária, a codificação, através de uma taxionomia controlada, pode incrementar a eficiência da comunicação eletrônica e o processamento das informações clínicas codificadas, melhorando a qualidade dos dados armazenados (Letrilliart, Viboud e

Boëlle, 2000; Tretiakov, Hunter, *et al.*, 2006). No entanto, observa-se que a codificação, quando feita, é geralmente realizada com o propósito exclusivo de cumprir os critérios definidos pelos mecanismos de reembolso, sem objetivar a correta identificação de um quadro clínico, de forma a apoiar a decisão e a recuperação de informações (Tretiakov, Hunter, *et al.*, 2006). Como instância do cenário apresentado, Vera e Martins (1994) analisaram a documentação clínica de uma amostra de hospitais do Rio de Janeiro e concluíram que a confiabilidade de diversos campos codificados na AIH (autorização de internação para o Sistema Único de Saúde - SUS) é insatisfatória. Nos "testes cegos" realizados, os índices de concordância para a codificação dos Serviços Auxiliares de Diagnósticos e Terapia - SADT, atingiram 51,6% para as descrições de Anatomia Patológica (mais freqüentes) e apenas 15,1% para Nebulização. A codificação do procedimento principal, causa da internação, no mesmo estudo, apresentou grau de concordância de 72%, considerando apenas os documentos devidamente preenchidos (abrangendo amostra de 79% dos documentos disponibilizados para a pesquisa, os 21% descartados não foram devidamente preenchidos pelo médico responsável). No caso do diagnóstico secundário, foi identificado o apontamento em apenas 1,9% dos casos.

Objetivando colocar em prova os dados apresentados por Vera e Martins (1994), pesquisa similar foi realizada nos hospitais da Aliança Saúde[2] em junho de 2009, envolvendo 8 profissionais médicos, 4 de cada hospital, com a análise de 300 AIH de média complexidade clínica, o grau de concordância levantado foi de 76%, percentual similar ao obtido pelas pesquisadoras, na mesma direção, o diagnóstico secundário é identificado em apenas 8% dos prontuários analisados.

Cabe ressaltar que a codificação parcial, incompleta ou ainda incorreta dos dados que compõem o registro clínico é geralmente ocasionada por falta de treinamento da equipe clínica, pela quantidade de códigos disponíveis (que no caso do CID ultrapassa 14000 possibilidades) e não normalizados ou, ainda, pela inexistência de um código específico para o quadro desejado (Porcheret, Hughes e Evans, 2004; Sanderson, Adams, *et al.*, 2004; Faulconer e Lusignan, 2004).

Neste contexto, o uso de texto livre, comumente formado ao longo do atendimento do paciente, seja na forma de anamneses, evoluções ou sumários de alta (narrativas clínicas), é elemento fundamental para o acompanhamento do quadro clínico do paciente, por complementar a codificação e, geralmente, por formar a fonte mais completa de análise das informações relacionadas à assistência.

---

[2] Hospital Cajuru, Hospital Santa Casa de Misericórdia, Hospital Nossa Senhora da Luz e Hospital Maternidade Alto Maracanã

Para o processamento computacional de texto livre, torna-se necessária a aplicação de técnicas de Processamento de Linguagem Natural – PLN. Como desafio adicional para o cenário das narrativas clínicas, destaca-se a prevalência de textos ‘sujos’, com grande quantidade de erros e baixíssimo formalismo estruturante (Ertmer e Ückert, 2005). Complementarmente, é necessário o mapeamento do conhecimento não estruturado (presente nas narrativas) para elementos formais ou semi-formais que permitam a manipulação (ex.: inferência e busca) do conhecimento representado. Uma estratégia possível é adoção de terminologias, como o CID10, disponível para o domínio médico, mas limitado para a representação de doenças, inviabilizando a plena representação do conteúdo documental e, por conseguinte, dificultando a recuperação e manipulação de informações. Outra solução possível é a adoção de ontologias e, no caso do domínio médico, a SNOMED CT (IHTSDO, 2009) apresenta-se como uma alternativa abrangente e viável.

Outro fator a ser considerado, na dimensão da adoção prática de qualquer estratégia a ser proposta, é a necessidade de produção de baixo impacto na rotina de trabalho dos profissionais de saúde. As técnicas de PLN, bem como as de manipulação de ontologias, são computacionalmente complexas e intensivas (Moghrabi, Moussa e Eid, 2002; Jeschke, Natho e Wilke, 2007). Neste sentido, a opção de mecanismos que possam assistir ao profissional médico durante o atendimento clínico, como apoio a decisão e a preparação da documentação exigem a adoção de técnicas de engenharia de software além das tradicionalmente utilizadas, com foco no paralelismo, robustez e cooperação, com foco na geração de resultado qualitativo no menor intervalo de tempo possível.

Como instância do quadro apresentado, nos hospitais da Aliança Saúde, o tempo médio despendido no preenchimento de uma evolução clínica, entre janeiro de 2009 e junho de 2009, para pacientes internados sem passagem por Unidade de Terapia Intensiva - UTI, é de 2,3 min e de 6 min para pacientes em tratamento em UTI. Para os atendimentos ambulatoriais, o atendimento clínico para pacientes do SUS demanda, aproximadamente, 12 min de atenção profissional.

Na dimensão técnica, em específico no domínio da engenharia de software, o processamento computacional de texto livre aplicado ao domínio da saúde demanda a adoção de estratégias que considerem a necessidade de paralelismo e escalonamento, neste sentido, os sistemas multiagente (tema enquadrado na área de Inteligência Artificial Distribuída - IAD) apresentam grande aderência, permitindo a construção de um ambiente altamente paralelizável, com simplificada incorporação de novas estratégias (Dawit e Davidsson, 2008).

No contexto apresentado, a literatura disponível apresenta casos em áreas correlatas a esta pesquisa, que serão explorados nos parágrafos que seguem.

Schulz e Mihov (2002) enumeram bases para o processamento de textos médicos, dando ênfase aos aspectos lingüísticos e na análise multi-lingual, fundamental quando da inter-relação de conteúdo disponível em múltiplas línguas. O trabalho, no entanto, não discute soluções quando do processamento de narrativas clínicas, caracterizado por textos 'sujos', também não discorre sobre a necessidade de pré-processamento de textos livres, de forma a tratar casos de uso de acrônimos, erros ortográficos e, ainda, da expansão de códigos utilizados nesses textos (como a codificação de exames através da adoção de alguma das terminologias disponíveis para este fim).

Weber-Jahnke e Price (2007) relacionam os principais requisitos de um sistema hospitalar, destacando aspectos de arquitetura, usabilidade e segurança da informação. Os autores destacam a necessidade da adoção de padrões de interoperabilidade, a relevância clínica para a codificação baseada na SNOMED CT (IHTSDO, 2009) e a importância no tratamento de textos livres no domínio médico, sem, no entanto, discorrerem sobre as estratégias para tanto.

No contexto de mapeamento de texto para estruturas computacionais, Hahn *et al.* (2004) apresentam estratégias para o mapeamento de resumos de artigos para o MESH, que é um sistema de entradas para catalogação de trabalhos científicos. O trabalho estabelece as bases para o mapeamento de resumos de textos científicos para uma terminologia, no entanto, foca apenas os aspectos de mapeamento, não enunciando estratégias para a adoção de uma ontologia e o aproveitamento das relações, características de uma ontologia de domínio.

Wong *et al.*(2007) apresenta estratégias para a manipulação de textos 'sujos', enfatizando textos disponíveis na Web, que apresentam similaridade construtiva com as narrativas clínicas, por não apresentarem elevado formalismo quando da escrita. Pelo enfoque generalista da publicação, o autor não discorre sobre o tratamento especial necessário quando da aplicação das estratégias em textos livres técnicos, caracterizados pela ampla adoção de acrônimos, siglas, valores numéricos, entre outros; desafio presente quando do tratamento de documentos clínicos.

Qamar e Rector (2007) destacam a importância para o mapeamento de estruturas documentais para ontologias, enfatizando a construção de modelos de uso (interfaces computacionais) e da relevância da SNOMED CT para este cenário, por representar uma terminologia que permite a documentação clínica nos mais diferentes universos de interpretação. Os autores ignoram o processamento de textos livres, focando suas pesquisas na construção de "dicionários de dados" que permitem a conversão de formulários codificados para um modelo comum, indexado a uma terminologia clínica, como a SNOMED CT.

Price *et al.* (2008) destacam que a inabilidade para a localização de respostas é um dos maiores obstáculos quando da manipulação de documentação médica em atendimentos clínicos. Respostas *just-in-time* são fundamentais para a melhora na qualidade do atendimento e maximização do tempo despendido. No trabalho, os autores propõem o mapeamento das principais questões clínicas obtidas através de estudo empírico. Essas questões foram mapeadas para descritores semânticos, assim como os documentos, através da comparação terminológica utilizando o UMLS e o MEDLINE, melhorando a precisão quando da busca de informações. A limitação do trabalho, uma das fronteiras explorada na presente tese, é a manipulação completa de texto livre e não apenas de descritores definidos e, também, a adoção de uma ontologia, que permite a expansão do universo de busca e manipulação através do uso das relações inerentes a uma ontologia.

No mesmo âmbito, Price *et al.* (2008), French *et al.* (2001) e Aronson (2001) enumeram possíveis estratégias para mapeamento de descritores em um vocabulário controlado, destacando que a estratégia melhora a eficiência da recuperação. A exemplo de Price, os autores não exploram a adoção de ontologias e suas relações semânticas como fronteira para um mapeamento mais eficaz.

Por fim, o uso de prontuário eletrônico poderá representar uma importante dimensão de mudança no cuidado à saúde, provendo dados para os mais diversos propósitos: atendimento ao paciente, suporte aos sistemas de tomada de decisão, pesquisa científica, validação e inter-relacionamento com *guidelines* e, mesmo, para o gerenciamento de unidades hospitalares (Moorman e Musen, 1999; Miettinen e Korhonen, 2008). A transição para registros eletrônicos demanda expressiva alteração na forma como as informações clínicas são expressas (Ammon, Hoffmann e Jakob, 2008), em especial, pela flexibilidade característica das narrativas clínicas. Segundo Lovis *et al.* (2005), o uso de texto livre continuará tendo importante papel para a documentação clínica, especialmente por melhorar a comunicação entre o paciente e corpo clínico.

Assim, processar narrativas clínicas, escritas em texto livre, textos estes com grande prevalência de erros, torna-se relevante para a extração de informação contida nessas narrativas. Complementarmente, mapear as informações para uma ontologia proverá mecanismos mais avançados que permitirão o manuseio e a recuperação das informações. Por isso, o uso da SNOMED CT apresenta-se como importante alternativa, representando uma terminologia clínica de ampla abrangência.

Então, motivado:

a) em geral, pela representação de conhecimento em domínio médico, em função dos diversos proponentes de conhecimento que o caracteriza;
b) pela necessidade de processos automatizados e inteligentes para o processamento de informações codificadas em narrativas clínicas;
c) pelo potencial que o domínio da inteligência artificial distribuída apresenta na modelagem de problemas complexos e distribuídos;
d) pela atenção que a comunidade científica tem dispensado no estudo e uso de ontologias de domínio; e
e) pela relevância do processamento de registros clínicos, como instância do uso e manipulação do conhecimento biomédico, desafio primeiro para as ciências médicas no século XXI (Pestotnik, 2000).

Na presente tese, discute-se e propõem-se novas metodologias e tecnologias para o processamento de registros clínicos, possibilitando meios para a recuperação dos mesmos através do mapeamento dos artefatos processados em uma ontologia de domínio.

1.2 Objetivos

Como objetivo geral, investiga-se e propõe-se heurísticas para processamento e mapeamento **de informação** codificada em registros clínicos, armazenados em meio eletrônico, baseadas em **ontologias e agentes** computacionais, para uma ontologia de domínio, para melhorar a indexação e recuperação de documentos.

No intuito de alcançar o objetivo geral, estratificam-se os seguintes objetivos específicos:

a) **discutir** a fundamentação teórica;
b) **definir** heurísticas e estratégias para o processamento de linguagem natural em narrativas clínicas;
c) **avaliar** experimentalmente as heurísticas propostas;
d) **desenvolver** estratégias para a normalização morfossemântica;
e) **mapear** narrativas clínicas para ontologias de domínio, através da aplicação das heurísticas avaliadas, disponibilizando, para tanto, componentes de software que sirvam como base para construção de ferramentas de apoio à decisão e recuperação de informação;

A partir da análise de narrativas clínicas reais e utilizando uma ontologia no domínio médico:

f) **criar** um repositório, o qual deverá acondicionar a ontologia em si, um vocabulário específico (incluindo jargões, acrônimos e demais características idiomáticas do cenário considerado) e registros clínicos;
g) **especializar** as técnicas desenvolvidas de forma a otimizar o mapeamento de informação;
h) **desenvolver** padrão para a validação de resultados; e
i) **validar** os resultados obtidos.

1.3 Justificativa

O aumento exponencial da disponibilidade dos dados confere significativa importância às técnicas de organização da informação. Técnicas que fazem parte de um conjunto de disciplinas que buscam melhorias no tratamento dos dados, atuando na sua seleção, processamento, recuperação e na disponibilidade dos mesmos.

Na área médica, os dados que descrevem um episódio clínico, englobam bases de dados e registros médicos textuais. Para o alcance do processamento universalizado desses dados, registrados segundo um conjunto de práticas, em um domínio específico, é necessária a adoção de uma estrutura que permita a conversão do domínio específico (termos e linguagem pertinentes ao documento) para um senso comum (mesmo que local). Neste sentido, dentre as técnicas disponíveis, destaca-se a adoção de ontologias de domínio (Bodenreider, Smith, *et al.*, 2007).

Para o processamento de registros médicos, mapeando-os para uma ontologia de domínio, através de componentes passíveis de tratamento por computador, há a necessidade de se recorrer a dois grupos de disciplinas:

a) as que auxiliam na compreensão do problema a ser estudado e modelado; e
b) as que contribuem com conceitos e metodologias que são suportadas através de recursos computacionais e permitem a implementação da solução adotada.

No primeiro grupo, ou seja, as que auxiliam na compreensão do problema, fundamentalmente, encontram-se: a engenharia de conhecimento, a lingüística computacional, a teoria geral de sistemas, o estudo de terminologias médicas, a engenharia de software e, na análise contextual dos artefatos processados, a medicina, nas suas mais diversas especialidades. Dentre as que fazem parte do segundo grupo, destaca-se a inteligência

artificial, mais especificamente a inteligência artificial distribuída no seu ramo de sistemas multiagente e o processamento de linguagem natural.

Com a finalidade de enquadrar o tema da tese nas disciplinas mencionadas, será primeiramente necessário enunciar o problema a ser estudado e modelado. Para tal, considere-se o seguinte cenário:

Dado o conjunto de artefatos documentais, especificamente registros médicos, os quais:

a) são realizados com o propósito de documentar a evolução clínica de um paciente;
b) são produzidos segundo conjunto de metodologias, registrados e armazenados através de técnicas disponibilizadas pelos Sistemas de Informação;
c) permitem a realização do registro clínico por diferentes profissionais, das mais variadas especialidades; e
d) consistem de documentos escritos em linguagem natural, produzidos sem rigor literário, em função das limitações temporais e de recursos, mesclando informações textuais, valores numéricos (por exemplo: temperatura) e informações estruturadas (por exemplo: código de procedimento), então, buscando o processamento dos registros médicos, de forma a inter-relacionar o conhecimento representado em uma ontologia de domínio, as seguintes ações fazem-se necessárias:
    i) pré-processamento dos registros clínicos, de forma a resolver as ambigüidades relacionadas ao uso de acrônimos, explicitação de siglas, delimitação de valores numéricos e correção ortográfica do texto de entrada;
    ii) identificação das frases nominais, obtidas através da aplicação de ferramentas de processamento de linguagem natural;
    iii) pós-processamento das frases nominais, com a aplicação de técnicas baseadas nas metodologias de recuperação de informações, para o mapeamento para ontologias de domínio; e
    iv) criação manual de um padrão de referência (objetivando a avaliação dos resultados obtidos).

Considerando o cenário descrito, o tema central da tese contempla a elaboração de uma abordagem baseada em inteligência artificial, que modele as tarefas de processamento de

registros clínicos como agentes computacionais (artefatos) e o mapeamento dos mesmos para uma ontologia de domínio como uma das capacidades desses agentes.

Para a consecução do cenário apontado, é fundamental considerar as formas complexas e idiossincráticas na geração de termos médicos. Diante dessas peculiaridades, os atuais métodos de recuperação e extração de informação, implementados, por exemplo, em máquinas de busca, baseados em comparação simples de palavras inteiras, sem um processamento lingüístico, mostram-se inadequados, pois produzem recuperações incompletas, imprecisas ou fora do âmbito desejado (Schulz e Hahn, 2002). Outros problemas decorrem da complexidade das ferramentas de busca e na tecnologia empregada para construção dessas ferramentas (Wille e Bruza, 1995). Também ocorre que as ferramentas agrupam as informações de diferentes maneiras, cada uma com seu modelo específico que atinge somente aqueles que conhecem o funcionamento interno, o que torna mais fácil o encontro de informações pertinentes (Iivonen, 1995).

Outra dimensão a ser considerada é a dificuldade ocasionada pela multidisciplinaridade e interdisciplinaridade da produção do histórico clínico do paciente. São diversos profissionais, com diferentes formações, descrevendo e registrando eventos sob a ótica de sua especialidade, utilizando acrônimos, tabelas de codificação, contrapondo terminologias controladas e informais..., enfim, diferentes terminologias complementares e paralelas que, quando combinadas, descrevem com maior exatidão o momento clínico do paciente (Bossen, 2006; Eichelberg, 2005; Skov, 2006).

Neste domínio, processar e recuperar registros clínicos, caracterizados pela multidisciplinaridade, em tempo hábil, passa a ser um desafio contínuo em instituições de saúde, especialmente em hospitais universitários, para a consecução de suas pesquisas, e ferramenta de apoio à decisão para os profissionais da área na busca da melhora contínua de processos (Lovis, Baud e Planche, 2005).

Para a construção de ferramentas de apoio à decisão, a identificação de bibliografia correlacionada ao registro clínico, é requisito funcional. Para tanto, é necessário considerar que muitos documentos relevantes estão disponíveis somente em línguas diferentes da língua nativa do usuário, prejudicando a eficiência da recuperação pretendida. O inglês, sendo a língua de referência, principalmente com a popularização da Internet, não é adequadamente dominado pelos usuários que, mesmo com maior ou menor capacidade de ler documentos no idioma estrangeiro, sentem dificuldades ao formular consultas adequadas nos sistemas de recuperação (Schulz e Mihov, 2002). Neste cenário, quatro são os desafios identificados (Pacheco, Stenzhorn, *et al.*, 2009): recuperação intra-lingual (no idioma nativo, incluindo jargão técnico), trans-lingual (entre línguas diferentes ou dentro de coleções de documentos

multilíngües), processamento de linguagem natural e mapeamento de frases chaves de narrativas clínicas para terminologias clínicas e ontologias de domínio. O processamento ocorre em textos contidos em diferentes repositórios de dados, contextos e aplicações.

A tarefa de mapeamento automático de textos para uma estrutura de representação de conhecimento (independente de qual seja) apresenta-se como um processo de grande relevância por ser pré-requisito para muitas aplicações e demandas incluindo: suporte à tomada de decisão (Fiszman e Haug, 2000; Friedman, Knirsch e Shagina, 1999), recuperação de informação (Hersh, Mailhot e Arnott-Smith, 2001), *text mining* (Gundersen, Haug e Pryor, 1996), classificação e categorização, sumarização de texto, resposta automática a questões, descoberta de conhecimento e outras relacionadas ao processamento de linguagem (Bashyam e Taira, 2007).

O presente trabalho apresenta-se de forma inédita, ao desenvolver, discutir e propor soluções aplicadas para o mapeamento de narrativas clínicas, pontuando estratégias para a manipulação das mesmas (tratando a complexidade inerente às narrativas: prevalência de acrônimos, a não observância das regras lingüísticas, etc) para uma ontologia de domínio (utilizando um normalizador morfosemântico, conjuntamente com estratégias de recuperação de informação e utilização de relações ontológicas).

1.4 Organização do documento

Com o objetivo de fornecer uma visão geral do conteúdo e da organização da tese, a Figura 1 apresenta de forma sucinta a estrutura da mesma.

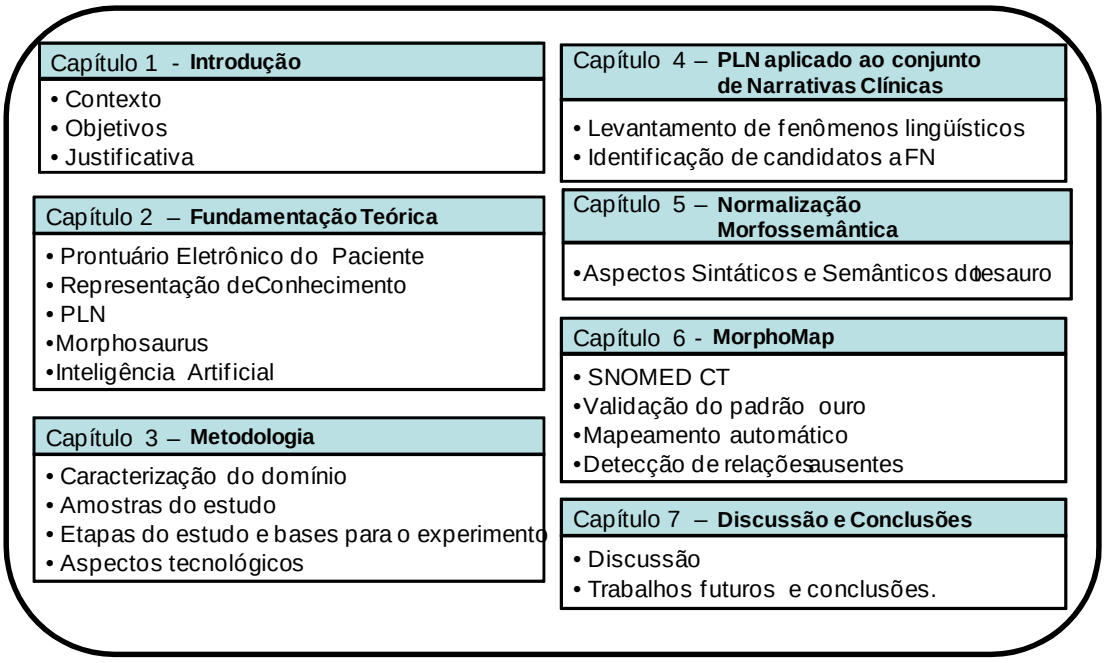


Figura 1 — Estrutura da Tese

O Capítulo 2 contém a primeira parte dos elementos metodológicos e conceituais que são aplicados na construção da presente pesquisa, passando por conceitos que fundamentam o tema, como a Inteligência Artificial, até metodologias que permitem a modelagem e enquadramento da solução, como Recuperação de Informação e Representação de Conhecimento, discorrendo também, sobre estruturas que fundamentam o domínio, como o Prontuário Eletrônico do Paciente, objeto de manipulação da presente, e o Processamento de Linguagem Natural.

No Capítulo 3 é apresentada a metodologia de desenvolvimento da tese.

O Capítulo 4 apresenta os elementos diferenciadores do Processamento de Linguagem Natural aplicado ao domínio das Narrativas Clínicas.

O Capítulo 5 apresenta os elementos da Normalização Morfossemântica, com ênfase nas diversas modificações e iniciativas realizadas objetivando a adequação do tesauro do sistema MorphoSaurus para a utilização do mesmo em um cenário real, como o da presente tese.

O Capítulo 6 contém as iniciativas desta pesquisa, apresentando os resultados alcançados na metodologia, em comparação com o padrão ouro constituído.

Por fim, no Capítulo 7 são apresentadas as considerações finais da presente tese e a proposição de trabalhos futuros, com base nas oportunidades científicas identificadas durante o desenvolvimento da presente tese.

Capítulo 2

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Introdução

Sayão (1996) destaca que a ciência é um processo que se apropria de maneira seletiva das contribuições de seus pesquisadores, tem um caráter cumulativo resultante de um corpo de conhecimento baseado em consenso, oferecendo à comunidade grupos disciplinares para a construção coletiva do conhecimento. Pode-se considerar "a ciência como um amplo sistema social, no qual, a partir dos conhecimentos consensuais fundamentados e materializados por meio da literatura técnico-científica, cumpre-se a função de disseminação do conhecimento" (Sayão, 1996).

Neste domínio, para a modelagem de estratégias de mapeamento de texto livre para estruturas de representação de conhecimento, através de componentes passíveis de tratamento computacional, necessário é recorrer a três grupos de disciplinas:

a) as que auxiliam na compreensão do problema a ser estudado e modelado;
b) as que contribuem com conceitos e metodologias, que são suportadas através de recursos computacionais e permitem o desenvolvimento da solução adotada; e
c) aquelas que estruturam os artefatos e o meio de ação.

No primeiro grupo, ou seja, as que auxiliam na compreensão do problema, fundamentalmente encontram-se: a gestão de conhecimento e lingüística. Dentre as que fazem parte do segundo grupo, destaca-se a Inteligência Artificial - IA, por propor estratégias eficientes para a solução de problemas complexos (George, 2001), especificamente no domínio dos Sistemas MultiAgente, por disponibilizar métodos e ferramentas para o

processamento distribuído (Adelinde e Weyns, 2009); a área de ontologias, por estruturar a representação de conhecimento (Cimiano, 2006); e o Processamento de Linguagem Natural, ao identificar as estratégias necessárias para a manipulação computacional de textos (Baeza-Yates e Ribeiro-Neto, 1999). No terceiro grupo, destacam-se a legislação e as metodologias relacionadas à produção de documentação clínica, especificamente aqueles relacionados ao prontuário eletrônico do paciente.

Devido ao tema proposto apresentar-se como sendo multi e interdisciplinar nos vetores de compreensão, modelagem e estruturação, é possível entender os cenários e as disciplinas que o compõem, de forma interconectada, em uma contínua realimentação, contribuindo para a expansão do conhecimento sobre o elemento em estudo, nas várias disciplinas que o suportam.

2.2 Prontuário Eletrônico do Paciente

Prontuário do paciente é o conjunto de documentos, padronizados e ordenados, destinado ao registro dos cuidados profissionais prestados durante a assistência médica. Sua finalidade é sistematizar informações necessárias para garantir a continuidade do atendimento profissional ao paciente, apoiar o ensino e a pesquisa ou oferecer suporte a qualquer demanda legal (Weaver, 2002).

Ao considerar que o prontuário é um documento do paciente, porém, a guarda do mesmo é de responsabilidade do hospital ou da unidade de saúde na qual o paciente foi assistido. Este documento deve estar sempre disponível para fins de continuidade ao atendimento, consulta ou resolução de pendências judiciais, conforme define resolução nº 1.331/89 do Conselho Federal de Medicina - CFM, o prontuário médico é “um documento de manutenção permanente pelos estabelecimentos de saúde”. Depois de decorrido prazo de dez anos, contados a partir da data do último registro de atendimento ao paciente, o prontuário pode ser substituído por outros métodos de registro, desde que capazes de assegurar a restauração das informações nele contidas de maneira integral.

Em complementação à resolução 1.331/89 do CFM, o estatuto da Criança e do Adolescente, pela lei nº 8.069/90 (CÓDIGO CIVIL, Lei nº 8.069, 1990), estabelece, pelo artigo 10, que hospitais e demais estabelecimentos de atenção à saúde de gestantes, sejam públicos ou particulares, são obrigados a manter o registro das atividades desenvolvidas, por meio de prontuários individuais, pelo prazo de pelo menos dezoito anos (CFM nº 2.539/1993).

Segundo a resolução CFM nº 1.605/2000, o médico deve fornecer cópia da ficha ou do prontuário médico desde que solicitado pelo paciente ou requisitado pelos Conselhos Federal ou Estadual de Medicina.

Klück e Guimarães (2002) destacam que se consideram permanentes os conjuntos de documentos de valor histórico, probatório e informativo que devem ser definitivamente preservados. Os documentos de valor permanente são considerados inalienáveis e imprescritíveis. Documentos de guarda permanente são aqueles que podem ser utilizados para fim de estudos e pesquisas médicas, científicas, sociais ou históricas, configurando, assim, um valor secundário ao documento.

Com base no contexto legal apresentado e considerando que a resolução CFM nº 1.331/89 não determina a forma de armazenamento do prontuário, destaca-se a necessidade de adoção de tecnologias de informação para a gestão desses documentos, dado o volume físico demandado para a sua guarda. Como instância do problema ora apresentado, dados dos Hospitais Cajuru e Santa Casa de Misericórdia de Curitiba levantados em 2007, destacam a necessidade adicional média mensal de 8m^3 para a guarda dos documentos relacionados às 600 internações médias destes hospitais e dos 30.000 consultas ambulatoriais realizadas[3].

Destaca-se a necessidade essencial, no contexto apresentado, de acesso eficiente e veloz a este conjunto de documentos; desta forma, a adoção de tecnologias de informação tem se mostrado um aliado importante nesta frente de trabalho (Dick, 1991).

Neste cenário, a adoção do prontuário eletrônico é estratégia central na maioria das unidades de saúde no mundo (Win e Susilo, 2006), não sendo diferente no Brasil.

Cabe ressaltar que, além da otimização no armazenamento, a adoção de tecnologias de informação permite a otimização de processos clínicos e administrativos (Dorileo, Ponciano e Costa, 2006), bem como cria uma importante fronteira de pesquisa, ao permitir a aplicação de técnicas de recuperação de informação, representação de conhecimento e processamento de linguagem natural nos documentos que compõem o prontuário (Win e Susilo, 2006).

Outro fator determinante para a adoção do prontuário eletrônico, em detrimento ao prontuário em papel, diz respeito à segurança e confidencialidade das informações codificadas nos diferentes documentos clínicos. Segundo a resolução CFM Nº 1.246/1998, é vetado aos profissionais de saúde a revelação, sem o consentimento do paciente, do conteúdo do

---

[3] Base de cálculo considerando que os prontuários, nos hospitais citados, são acondicionados a uma pasta de 23 por 32 cm, sendo que as internações, com média geral de 5 dias, geraram documentação (laudos de exame, prescrições, evoluções, entre outros) com altura aproximada de 3,5 cm (considerando que os documentos são acondicionados em uma pilha contígua), considerando que o número médio combinado e mensal de internação nas unidades referenciadas no ano de 2007 foi de aproximadamente 600 pacientes, a necessidade mensal de área para acondicionamento foi, aproximadamente, igual a 6,62 m^3.
Os atendimentos ambulatoriais, para os hospitais citados, sob a mesma ótica, produziram altura média de 3 mm. Para os 30000 atendimentos médios foram necessários 1,54 m^3 mensais, para o ano de 2007.

prontuário ou ficha médica do paciente. No caso da instrução de um processo criminal, a apresentação do conteúdo do prontuário ou da ficha médica pode ser solicitada por uma autoridade judiciária competente. Neste caso, o médico deve disponibilizar os documentos ao perito nomeado pelo juiz, para que neles seja realizada perícia restrita aos fatos em questionamento.

2.3 Conceitos ligados à Representação de Conhecimento

Todo problema a ser resolvido requer um conjunto de conhecimentos que dê suporte à sua solução. Para se utilizar este conjunto de conhecimentos em uma máquina, é necessário selecionar uma maneira de representá-lo. Representação do Conhecimento, portanto, é o modo de como os conhecimentos sobre o mundo podem ser representados no computador e que tipos de raciocínio podem ser aplicados na base conceitual constituída (Michalsky e Kodratoff, 1990; Maryam e Leidner, 2001; Puustjärvi, 2009).

A representação do conhecimento é um assunto multidisciplinar englobando (Sowa, 1999):

a) **Lógica**: a qual fornece a estrutura formal e regras de inferência;
b) **Ontologia**: que define as estruturas contempladas no domínio de aplicação; e
c) **Computação**: que estabelece as rotinas para o processamento do conhecimento representado.

Dentre as técnicas para representação de conhecimento, a de mapas conceituais é uma das mais antigas, tendo sido apresentada por Novak (1977), é baseada na teoria de David Ausubel, que destacou a importância do conhecimento prévio como sendo a base para engendrar novos conceitos, baseada na teoria construtivista.

Novak (1977) destaca que "aprendizagem significativa envolve a assimilação de novos conceitos e proposição em estruturas cognitivas existentes".

Os mapas conceituais podem ser descritos sob três perspectivas conforme o nível de análise a ser considerado (Gaines e Mildred, 1995):

a) **abstrata**: os mapas conceituais podem ser tratados como hiper-grafos ordenados, constituídos por nós ligados por arcos, sendo que cada nó é representado por um identificador e um conteúdo, já as ligações entre nós podem ser direcionadas ou não, representados visualmente por linhas entre os nós, com ou sem setas nas extremidades;
b) **visualização**: os mapas conceituais podem ser tratados como diagramas, representados através do uso de signos. Cada tipo de nó pode determinar (ou ser determinado) pela

forma ou cor, enquanto as ligações podem ser identificadas pela espessura da linha, cor ou outras formas de representação; e

c) **conversação**: os mapas conceituais podem ser considerados como uma forma de representação e comunicação do conhecimento através de linguagens visuais.

Um mapa conceitual parte de um conhecimento inicial, bem generalizado, a partir disso, constrói-se uma rede de conhecimentos que, de uma maneira geral, especializam-se ou relacionam-se com o conhecimento inicial.

Já as redes semânticas são utilizadas na definição de um amplo conjunto heterogêneo de soluções. De uma forma geral, a característica comum a estes sistemas é a notação: uma rede semântica é constituída de um conjunto de nós conectada por um conjunto de arcos. Os nós, de forma geral, representam objetos e os arcos relações entre esses objetos.

A formalização apresentada foi proposta por Quillian (1968), por meio de um modelo computacional da memória humana denominado de memória semântica.

Duas grandes contribuições para a linha de pesquisa de Redes Semânticas adveio de Minsky, ao propor o formalismo de *frames* (Minsky, 1975) e Woods ao analisar o significado dos arcos nas redes semânticas (Woods, 1975).

Minsky (1975) apresentou a estrutura dos *Frames*[4] como a base para o entendimento da percepção visual, diálogos de linguagem natural e outros comportamentos complexos.

Os *Frames* possibilitam a representação das estruturas internas dos objetos, mantendo a possibilidade de representar herança. O método de *frames* também está ligado à origem das idéias que levaram às linguagens de Programação Orientada a Objetos - POO.

Uma grande inovação diz respeito à introdução de nós com estrutura interna. Tal estrutura interna foi possível de ser idealizada graças ao mecanismo representacional proporcionado pelo atributo, que representa o local onde uma parcela do conhecimento é armazenada.

Sendo assim, um *frame* consiste em um conjunto de atributos que, através de seus valores, descrevem as características de um objeto que representa alguma entidade do mundo. Os atributos interligam-se a outros *frames* criando uma rede hierárquica dependente, atrelando a estrutura como um todo.

Como o *frame* possui a capacidade de herança de propriedades, pode-se especificar uma classe de objetos que é instanciada em uma subclasse, sendo que esta usufrui de todas as propriedades da sua super-classe.

---

[4] A literatura apresenta também a designação classe, entidade ou objeto; para um *frame.*

Outro conceito importante é o raciocínio guiado por expectativas, dado que um *frame* contém atributos, e estes podem ter valores típicos. Ao instanciar um *frame* para que corresponda a uma dada situação, o processo de raciocínio deve preencher os valores dos atributos com as informações disponíveis na descrição da situação. O fato de que o processo de raciocínio conhece as informações necessárias, e o caso destas informações estarem disponíveis é preponderante para a eficiência do reconhecimento de uma situação complexa (Costa e Bittencourt, 2000).

Um *frame* deve apresentar uma via responsável pela inicialização dos valores do *frame*, ou seja, a definição inicial de seu estado interno; uma função[5] de leitura, responsável pela captura de valores do *frame* e uma função de escrita responsável pela alteração do estado interno do *frame*.

As estratégias de Frames ou de Redes Semânticas são inadequadas para o tratamento de cenários de incertezas, nesta direção, as redes bayesianas fundamentam tal tratamento através do uso das evidências disponíveis (informação sobre o estado atual de uma variável aleatória) para refinar as estimativas das probabilidades associadas às demais variáveis de interesse. No entanto, as relações probabilísticas entre variáveis aleatórias são relacionamentos dinâmicos que se alteram quando evidências se tornam disponíveis. As redes bayesianas refletem esse comportamento dinâmico e fornecem o arcabouço teórico apropriado para se tomar decisões, com base em probabilidade, em ambientes com incerteza (Robert, 2007).

Essencialmente, a Teoria Bayesiana proporciona um formalismo matemático para explicitar a alteração de crenças atuais a partir de novas evidências, permitindo a combinação de novos fenômenos com conhecimentos pré-existentes.

O canônico exemplo consiste em imaginar que um recém-nascido precoce observa seu primeiro pôr-do-sol, e deseja saber se o sol irá nascer novamente ou não. Ele designa semelhantes probabilidades anteriores para ambos os acontecimentos possíveis, e representa isso através da colocação de uma bola branca e uma preta dentro de uma sacola. No dia seguinte, quando o sol nasce, a criança coloca outra bola branca na sacola. A possibilidade de que uma bola retirada aleatoriamente seja branca, portanto, aumentou de metade para dois terços. Após o nascer do sol no dia seguinte, a criança adiciona outra bola branca e a probabilidade (e o nível de crença) vai de dois terços para três quartos, e assim por diante.

[5] O termo função é o que melhor se emprega na situação. Função é um termo muito utilizado na área de informática e designa um procedimento capaz de retornar e receber valores.

Gradualmente, a crença inicial de que o sol pode nascer cada manhã é modificada para se tornar uma quase certeza que o sol irá sempre nascer (Murphy, 1988).

Robert (2007) define que o conhecimento empírico invariante (de um especialista) é codificado através da probabilidade condicional e a crença no evento conjunto A e C, se necessário, é estimada pela regra fundamental, sendo P(A|C) = p é a probabilidade condicional do evento A, dado que o evento C é verdadeiro e que as outras informações conhecidas são irrelevantes para A. Se P(A|B) = P(A), se diz que A e B são independentes. Se P(A|B,C) = P(A|C), A e B são ditas condicionalmente independentes, dado C. A regra fundamental para o cálculo de probabilidades é P(A|B)P(B) = P(A,B) ou a forma alternativa que explicita o contexto condicionante C, dada por P(A|B,C)P(B|C) = P(A,B|C). A regra da cadeia é dada por P(E1,E2,…,En) = P(En |En-1,…,E1) … P(E2 | E1)P(Ei).

Por fim, “usando o conceito da d-separação, Pearl provou que se por construção pa(Xi) $\subseteq$ {X1,…,Xi-1}; então, em uma rede bayesiana P(xi |xi-1,…,x1) = P(xi|pa(Xi). Logo, não é necessário representar todas as probabilidades dos eventos possíveis, resultando em um problema tratável em computador” (Robert, 2007).

Com a popularização das estratégias de mapeamento de conhecimento, identificou-se a necessidade de inter-relação entre as estratégias apresentadas no presente capítulo com o arcabouço conceitual da Filosofia. Populariza-se assim, na década de 1990, o uso de ontologias, nas mais diversas áreas do conhecimento, investigada por várias comunidades de pesquisa de IA, como as áreas de engenharia do conhecimento, processamento de linguagem natural e representação do conhecimento. Atualmente outras áreas têm também demonstrado interesse pelo uso de ontologias, como na integração entre domínios, recuperação de informação e gerenciamento do conhecimento (Alonso-Calvo e Maojo, 2007; Tempich e Staab, 2006).

Para a introdução do tema, inicialmente, faz-se necessário distinguir entre os termos “Ontologia” e “ontologia”: quando capitulada, o termo refere-se à disciplina filosófica, quando não, a um artefato (Guarino, 1998).

Segundo Kusniercyk (2006), o termo ontologia é claramente afetado por múltiplas interpretações inconsistentes e, assim, gerando expectativas irreais a respeito do que as ontologias podem alcançar (Stenzhorn, Schulz e Smith, 2008), demandando, assim, a prévia explanação de seu significado pretendido.

Mas o que é uma ontologia? No domínio da filosofia ontologia “é o estudo do que existe” (HOFWEBER, 2004). Muitas das clássicas questões filosóficas são problemas em ontologia, como a questão da existência, ou não, de Deus.

No domínio das ciências exatas, Mealey (1967) problematiza situações inerentes ao processamento de dados, ao enunciar o problema, contextualiza a aplicação do conceito de ontologia, antes inerente à Filosofia, no domínio da Ciência da Computação:

> "No mundo real as idéias existem na mente dos indivíduos, os símbolos existem em meios materiais ou em outras formas de armazenamento. Na verdade, é possível afirmar que os dados são fragmentos de uma teoria do mundo real e o processamento desses dados é a representação dos fragmentos das teorias relacionadas. Não existe uma forma única de representar um determinado dado, como é possível observar na representação do número cinco":
>
> **V $(101)_2$ 5_8 5 0.5E01**
>
> É possível identificar que todas as notações representam a mesma idéia, mas de diferentes maneiras".

Inúmeras são as definições para o termo ontologia. Algumas são apresentadas na seqüência.

Segundo Gruber (1993), uma ontologia define (ou especifica) os conceitos, relações, e outras distinções relevantes para a modelagem de um domínio. A especificação assume a forma das definições de terminologia representacional (classes, relações, e assim por diante), que dão significado aos termos e restrições formais para sua utilização coerente.

Huhns e Singh (Huhns e Singh, 1997) definem ontologia como "uma representação do conhecimento de algum domínio, que é deixada disponível para todos os outros componentes em um sistema de informação".

Weiss (1999) descreve ontologia como uma "especificação dos objetos, conceitos e relacionamentos na área de interesse. Uma ontologia é mais que uma taxonomia de classes (ou tipos), a ontologia deve descrever os relacionamentos".

Já Noy e McGuinness (2001) definem ontologia como a "descrição explícita formal de conceitos em um domínio do discurso (classes), propriedades de cada conceito descrevendo várias características e atributos dos conceitos (*slots*, chamados algumas vezes de papéis ou propriedades), restrições sobre *slots* (facetas, algumas vezes chamado de restrições do papel)".

Schulz & Johansson (2007) deliberadamente evitam o uso de termos ambíguos como "conceito" e "conhecimento" na sua tentativa de definir ontologia com relação à realidade como "um artefato de representação cujas unidades designam classes ou universos e suas inter-relações".

Em suma, ontologia é uma representação formal de entidades que podem ser instanciadas (conforme os autores chamados de conceitos, classes[6], tipos, universais) características e relacionamentos em um domínio específico, permitindo a interpretação comum em uma determinada área, propiciando, principalmente, o reuso de conhecimento.

O desenvolvimento de uma ontologia passa pela definição de classes na ontologia, organização das classes em uma hierarquia taxonômica e definição de propriedades no respectivo domínio (Noy e McGuinness, 2001).

De maneira geral, uma ontologia é desenvolvida para possibilitar o compartilhamento de conhecimento entre aplicações distintas, auxiliar a compreensão de alguma área do conhecimento e como um ferramental para a obtenção de consenso em alguma área do conhecimento (Gruber, 1993).

Gruber (1993) descreve os requisitos para a criação de ontologias, procurando maximizar o compartilhamento de conhecimento e a interoperabilidade entre programas:

a) **clareza**: uma ontologia deve efetivamente comunicar o significado pretendido dos termos definidos. Definições são objetivas. Enquanto a motivação para definir um conceito surge a partir de situações sociais ou requisitos computacionais, a definição deve ser independente do contexto social ou computacional;
b) **coerência**: uma ontologia deve ser coerente, isto é, ela deve permitir inferências que sejam consistentes com as definições. Coerência também deve ser aplicada para conceitos que são definidos informalmente, tal como aqueles descritos em documentação de linguagem natural e exemplos. Uma ontologia está incorreta quando uma sentença que pode ser inferida a partir de axiomas contradiz uma definição ou exemplo dado informalmente;
c) **extensibilidade**: uma ontologia deve permitir que novos termos possam ser definidos para usos especiais baseados no vocabulário existente, de maneira que não seja requerida a revisão das definições previamente existentes;
d) **mínimo compromisso com implementação**: a conceituação deve ser especificada a um nível de conhecimento sem depender de uma codificação particular a nível simbólico; e
e) **compromisso ontológico mínimo**: uma ontologia deve requerer o compromisso ontológico mínimo suficiente para dar suporte às atividades de compartilhamento de conhecimento desejadas.

[6] Sendo o termo “classe” o mais neutro, este será usado neste texto.

2.4 Processamento de Linguagem Natural

O PLN é uma subárea da IA e Lingüística, tendo como objeto a extração de informações computáveis a partir de textos. Os algoritmos de PLN demandam intenso processamento computacional (Jones e Somers, 1998), principalmente pelas características da linguagem natural, entre essas:

a) rica e elaborada e ao mesmo tempo vaga e ambígua;
b) os significados dos termos são, ao mesmo tempo, independentes e associados a outros termos, e
c) há inúmeras formas de se dizer a mesma coisa.

Muitos são os níveis de tratamento do PLN. Na Figura 2, apresenta-se parte da divisão proposta em Marcus (1980).

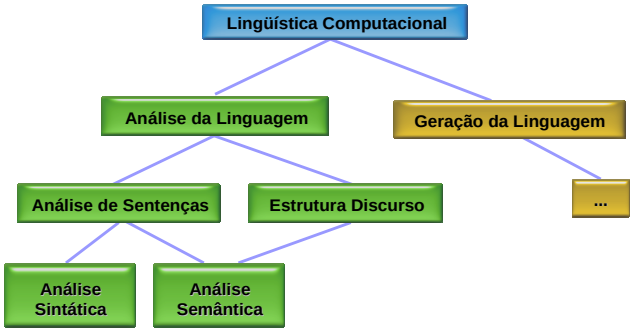


Figura 2 — Níveis de tratamento da PLN

Tradicionalmente, a primeira etapa do processamento da linguagem natural é a delimitação das sentenças e a identificação dos *tokens*. Posteriormente, as estruturas associadas a cada item, tais como: gênero e número para substantivos, ou pessoa, número, modo e tempo, para os verbos (Morton, 2006). Isso, porém, requer que as palavras possam ser identificadas mediante um léxico que também implementa o conhecimento da formação das palavras.

A saída do analisador léxico-morfológico é a entrada para o etiquetador gramatical (ou *POS (part of speech) tagger*), que é responsável pela etiquetagem, para cada item lexical, da categoria a que este item pertence. A etiquetagem é o "processo de demarcação de um marcador de classe gramatical (ou outro marcador ou 'etiqueta' de interesse) a cada palavra, num *corpus*" (Jurafsky e Martin, 2000).

Dois são os modelos fundamentais para os etiquetadores: baseado em regras e estocástico. No primeiro, regras são definidas de forma a permitir a identificação da categoria de um item lexical. A desvantagem deste mecanismo é que novas regras são manualmente adicionadas quando da identificação de novas situações. No caso do modelo estocástico, a

estratégia de ação é iniciada pelo treinamento baseado em um *corpus* previamente marcado, através do cálculo de probabilidade que um determinado item terá para cada etiqueta disponível (Jurafsky e Martin, 2000).

O analisador sintático contempla a etapa seguinte no PLN, objetivando o reconhecimento de uma seqüência de palavras como uma seqüência válida, ou não, da língua considerada. A análise sintática completa é dificultada pelo surgimento de um alto número de ambigüidades semânticas, assim como pela complexidade dos algoritmos. Na prática, o "*parser*" é um processo que analisa uma seqüência de entrada objetivando a identificação da estrutura gramatical segundo um determinado formalismo. Procura-se, neste cenário, chegar pelo menos a uma delimitação adequada de frases ou até entidades menores (*chunks*).

De forma complementar ao analisador sintático, Morton (2006) destaca o detector de frases nominais, que são aquelas que encerram em si significado completo, estático e independente.

2.4.1. Recuperação de Informação

A Recuperação de Informação é uma área de estudos que tem como meta o desenvolvimento de técnicas para busca de informações contidas em documentos ou a recuperação de documentos em si. Este é um tema multidisciplinar que envolve ciência de computação, semiótica, lingüística e biblioteconomia (Baeza-Yates e Ribeiro-Neto, 1999). Na maioria dos casos, a informação contida nesses documentos não é estruturada e não pode ser expressa formalmente.

Comumente as técnicas de recuperação de informação (Baeza-Yates e Ribeiro-Neto, 1999) são disponibilizadas sob a forma de mecanismos de busca e são implementadas de acordo com algum modelo de recuperação de informação (modelo booleano, modelo espaço vetorial, booleano estendido, difuso, probabilístico,...), sendo que o modelo espaço vetorial é o mais comumente aplicado.

Os mecanismos de busca (Das, Gunopulos e Koudas, 2006; Gomes, 2006; Marchiori, 1997; McBryan, 1994; Page, Brin, *et al.*, 1998) tradicionalmente exploram grandes sistemas de base de dados, construindo índices automatizados. Quando envolve o processamento de hipertexto, como a Web, recorre-se a um sistema de *Crawler* (Moshchuk e Bragin, 2006) para varrer a totalidade de documentos. Em essência, o artefato de busca dispara um conjunto de processos (*threads*, agentes distribuídos, etc)[7], que estabelecem a conexão com fontes de

[7] Conforme **Figura 3** item A

documentos[8], *a posteriori* os documentos são processados (sumarização, classificação, agrupamento, cálculo de similaridade), num processo conhecido como indexação (Baeza-Yates e Ribeiro-Neto, 1999)

O processo de *Crawler*, **Figura 3**, consiste no processamento de um determinado documento e a recuperação através dos *hyperlinks* configurados, de outros documentos relacionados (Moshchuk e Bragin, 2006). Num processo subseqüente, os documentos são avaliados e classificados, permitindo a posterior recuperação por parte do usuário.

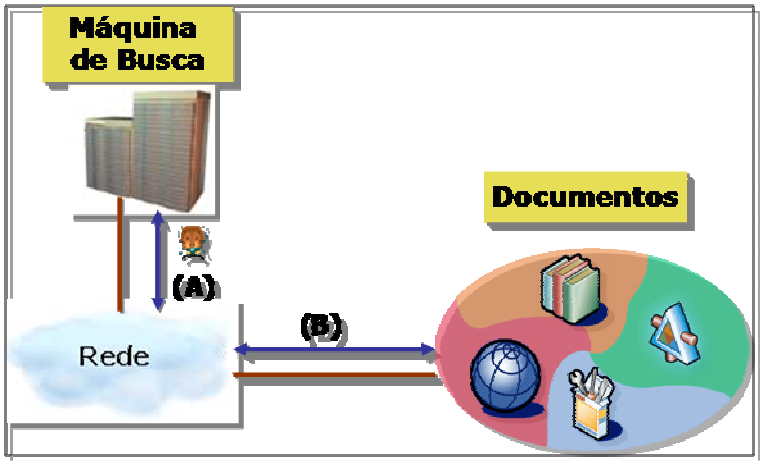


Figura 3 — **Indexação de documentos em máquinas de busca**

Em uma etapa posterior, os usuários, por intermédio de uma interface com o servidor, efetuam buscas estabelecendo um conjunto de critérios (**Figura 4**(A). O servidor, em resposta à consulta realizada, retorna um conjunto de referências para os documentos enquadrados nos critérios de busca (**Figura 4** (B).

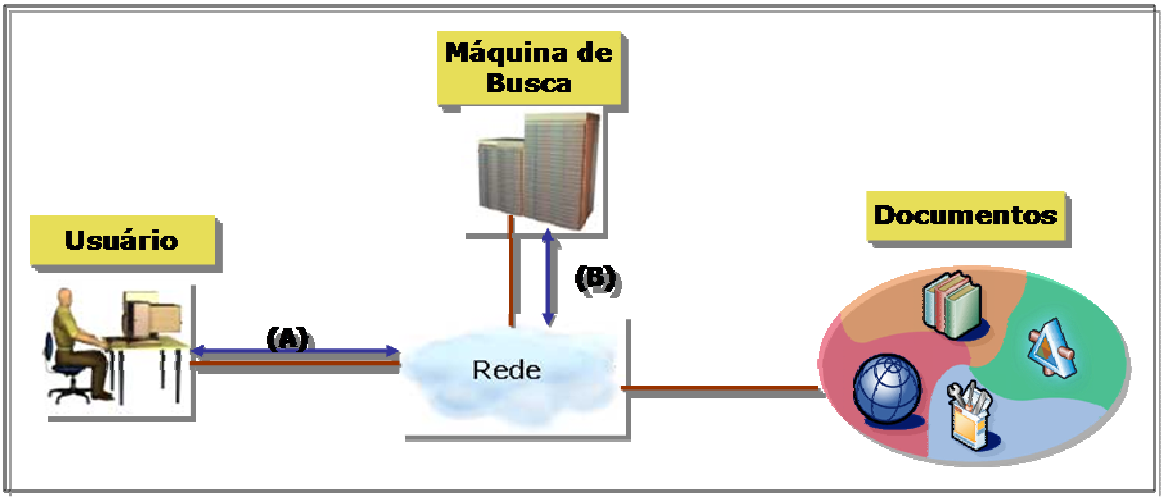


Figura 4 — Busca de documentos em máquinas de busca

[8] **Figura 3** item B

O aumento do repositório de documentos, propiciada, em grande parte, pela popularização da Tecnologia da Informação, gera um complicador adicional para o processo de indexação: a diversidade de fontes exige algoritmos e formas de indexação mais complexa e precisa, já que o resultado de uma busca passa a considerar um extenso domínio. Para tanto, inúmeros sistemas de recuperação de informação foram propostos, entre os quais, os sistemas de busca e de meta-busca que, em geral, utilizam técnicas de recuperação de informação baseadas no modelo vetorial proposto por Salton ou em uma de suas extensões[9] (Baeza-Yates e Ribeiro-Neto, 1999; Page, Brin, *et al.*, 1998). De modo geral, os sistemas de busca recuperam informações através de uma consulta baseada em palavras-chave. No caso citado, os documentos disponíveis na internet são processados de forma automatizada e alimentam a base de busca do sistema considerado. Sendo assim, quando da execução de uma consulta, os documentos que contenham as palavras-chave desejadas são recuperados, independentemente do domínio (natureza) do documento.

Outra solução adotada, de forma a delimitar o domínio, são os diretórios (Baeza-Yates e Ribeiro-Neto, 1999), utilizados em sistemas como o Altavista e Yahoo. Esses sistemas, adicionalmente, classificam os documentos em *frames* de conhecimento hierárquicos, permitindo a navegação a partir de um domínio amplo para um subdomínio mais específico. A limitação desta técnica de representação reside no processo manual de classificação, visto representar uma atividade de grandes proporções e na quantidade de documentos recuperados durante a navegação *top-down*. No mesmo domínio, Nascimento (2003) apresenta os portais especializados em certa área do saber, permitindo a especialização e a conseqüente limitação do domínio de interesse, ou ainda, o ambiente de bibliotecas digitais (Baeza-Yates e Ribeiro-Neto, 1999), anterior, inclusive, ao advento da Internet, tais sistemas, como o MEDLINE (www.ncbi.nlm.nih.gov), a BIREME (www.bireme.br) e o SciELO (http://www.scielo.org), são voltados para um público de uma determinada área (saúde, por exemplo) e não apenas para uma especialidade (exemplo: câncer).

O modelo espaço-vetorial de indexação foi desenvolvido por Gerard Salton, para utilização em um sistema de recuperação de informação denominado SMART (Baeza-Yates e Ribeiro-Neto, 1999; Salton e Buckley, 1987). Embora a proposta do modelo vetorial tenha surgido em 1968 como solução de problemas de busca, esse modelo é utilizado até hoje por ter bom desempenho na recuperação dos dados, visto que leva em consideração o casamento

[9] O uso do modelo vetorial de Salton é decorrente da relativa simplicidade de implementação e da eficiência quando aplicado a coleções genéricas de documentos (Baeza-Yates e Ribeiro-Neto, 1999).

parcial e a proximidade dos documentos em relação aos termos da consulta (Hersh e Buckley, 1994).

O modelo faz uso de uma lista invertida contendo o vocabulário do documento processado, onde a cada termo é atribuída um peso e a lista de documentos que apresentam o termo com sua respectiva freqüência.

Por intermédio do cálculo e comparação da similaridade entre a consulta submetida e os diversos documentos da coleção, é possível representar graficamente a relação de similaridade entre eles, obtendo-se, assim, as melhores respostas para a pesquisa.

Cada elemento do vetor de termos é considerado uma coordenada dimensional, podendo ser disposto em um espaço euclidiano de *n* dimensões (onde n é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso.

Existem várias formas de cálculo do peso do termo no documento. Entre os mais utilizados, encontra-se o modelo de Salton e Buckey (Baeza-Yates e Ribeiro-Neto, 1999), que procura balancear características em comum nos documentos (*intra-document : TF term frequency*) e características de distinção entre eles (*inter-document: IDF inverse document frequency*, Equação 1).

$$IDF(t) = \log\left(\frac{N}{n_t}\right) \qquad \text{Equação 1}$$

Onde N corresponde ao número total de documentos de uma coleção e n_t o número de documentos em que a palavra t ocorreu.

Para determinar as coordenadas do vetor do documento d no eixo t, utiliza-se a Equação 2:

$$w(d,t) = tf(d,t) * IDF(t) \qquad \text{Equação 2}$$

As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos com os mesmos termos ocupam uma mesma região do espaço. A consulta, de forma similar, também é representada por um vetor. Neste sentido, os vetores dos documentos podem ser comparados com o vetor de consulta, permitindo a obtenção do grau de similaridade entre os elementos considerados. Os documentos com maior similaridade (mais próximos no espaço) no domínio da consulta submetida são devolvidos, a similaridade é calculada com base na equação detalhada no Apêndice A.

2.4.2. Similaridade normalizada de Levenshtein

A similaridade normalizada de Levenshtein é baseada no algoritmo da distância generalizada de Levenshtein, que calcula o custo de transformação de um *token* em outro, analisando, geralmente, um conjunto de operações, como a inserção, remoção ou a substituição de um símbolo por outro (Sankoff e Kruskal, 1983).

Formalmente, a distância generalizada de Levenshtein é definida com base no alfabeto $\sum$, sendo $\sum^*$ um conjunto de elementos sobre $\sum$. O elemento $X \in \sum^*$ é denotado como $X=x_1x_2 \dots x_n$, onde x_i é o i-ésimo símbolo de *X*. $X_{i..j}$ é referido como os elementos de *X* que incluem os símbolos de x_i até x_j, com $1 \leq i \leq j \leq n$, o comprimento da cadeia é definido como $|X_{i..j}| = j - i + 1$ e o elemento nulo como $\lambda(|\lambda| = 0)$, para *i*>*j*.

Uma operação elementar de edição é definida pelo par $(a, b) \neq (\lambda, \lambda)$, também podendo ser denotado como $a \rightarrow b$, sendo que tanto *a* quanto *b* são elementos com comprimento 0 ou 1. As formas $\lambda \rightarrow a$, $a \rightarrow b$ e $b \rightarrow \lambda$ representam, respectivamente, operações de inserção, substituição e remoção. $T_{X,Y} = T_1T_2 \dots T_l$ é utilizado para denotar uma transformação de edição de *X* para *Y*, que representa a seqüência de operações elementares de edição para transformar *X* para *Y*. Por fim, se uma função γ, aplicada a $a \rightarrow b$, produzir um número real não negativo, o 'peso' de uma transformação de edição $T_{X,Y}$ pode ser computada como $[\![\gamma(T]\!]_{X,Y}) = \sum_{i=1}^{l} \gamma(T_i)$.

Dado $X,Y \in \sum^*$, o GLD é definido na Equação 3.

$$GLD(X,Y) = \min\{[\![\gamma(T]\!]_{X,Y})\} \qquad \text{Equação 3}$$

A distância normalizada de Levenshtein apresenta um desempenho melhor que o GLD quando da aplicação em cenários com muita variabilidade, gerando um menor número de falsos positivos, característica comum em aplicações de correção ortográfica (Yujian e Bo, 2007), sendo definido pela Equação 4.

$$d_{N-GLD}(X,Y) = \frac{2GLD(X,Y)}{\alpha * (|X| + |Y|) + GLD(X,Y)} = \frac{GLS(X,X) + GLS(Y,Y) - 2 * GLS(X,Y)}{GLS(X,X) + GLS(Y,Y) - GLS(X,Y)} \qquad \text{Equação 4}$$

Sendo que $\alpha = max\{\gamma(\lambda \square(\rightarrow \perp)\ a), \gamma(a \square(\rightarrow \perp)\ b), \gamma(\ b \square(\rightarrow \perp)\ \lambda), a, b \in \Sigma \equiv \}$ e a similaridade generalizada de Levenshtein, também conhecida como *Generalized Levenshtein Similarity* (GLS), entre *X* e *Y* é definida como $GLS(X,Y) = \frac{\alpha * (|X| + |Y|) - GLD(X,Y)}{2}$.

2.4.3. Etiquetagem morfológica

Dois são os modelos mais utilizados nos algoritmos para etiquetagem morfológica: os estocásticos e os baseados em regras. Algoritmos baseados em regras fazem uso de 'bases de regra' no processo de identificação da categoria de um dado item lexical. Normalmente, novas regras são agregadas à base no momento que novos contextos de uso são identificados. Já os algoritmos que são construídos com base em métodos estocásticos, determinam as etiquetas através do cálculo de probabilidade de um determinado item gramatical, em um contexto, contra uma etiqueta conhecida. As probabilidades são determinadas com base em *corpus* lingüístico anotado de treinamento e são mais adequadas quando da aplicação em contextos de sublinguagens específicas (Manning e Schuetze, 1999; Jackson e Moulinier, 2002).

No modelo estocástico, as seqüências de etiquetas em um texto são entendidas como uma cadeia (de segunda ordem) de Markov (Manning e Schuetze, 1999). Uma cadeia de Markov tem duas propriedades:

1) **horizonte limitado**, na qual a etiqueta de uma dada palavra depende apenas e tão somente da etiqueta anterior; e
2) **invariância temporal,** na qual uma etiqueta determinada não é modificada por observações futuras.

Instanciando as propriedades no domínio da etiquetagem morfológica, é possível assumir que a etiqueta de uma dada palavra depende somente da etiqueta imediatamente anterior (horizonte limitado) e que esta dependência não se altera ao longo do tempo (invariância temporal). Manning e Schuetze (1999) citam, como exemplo: "se um verbo infinitivo tem a probabilidade de 0,2 de ocorrer depois de um pronome, no início de uma sentença, então esta probabilidade não irá se alterar quando da etiquetagem do restante da sentença".

O estimador de máxima verossimilhança, t^k, determinado com base na análise do *corpus* manualmente anotado, é calculado com base no conjunto de elementos apresentados na Tabela 2.

Tabela 2 — Convenções de notação para o processo de etiquetagem morfológica proposto por **(Charniak, Hendrickson e Jacobson, 1993)**

w_i	palavra (“w*ord*”) na posição *i* do *corpus*
t_i	etiqueta (“*tag*”) de w_i
$w_{i,i+m}$	palavras que ocorrem da posição *i* até *i+m*
$t_{i,i+m}$	etiquetas $t_i ... t_{i+m}$ para $w_i ... w_{i+m}$
w^l	a l-ésima palavra no *corpus*
t^l	j-ésima etiqueta do conjunto domínio de etiquetas
$C(w^l)$	número de ocorrências de w^l no *corpus* de treinamento
$C(t^j)$	número de ocorrências de t^j no *corpus* de treinamento
$C(t^j, t^k)$	número de ocorrências de t^j seguida por t^k
T	número de etiquetas no conjunto domínio de etiquetas
W	número de palavras no *corpus* de treinamento
n	tamanho de sentença

t^j é estimado através das freqüências relativas das diferentes etiquetas com base em uma determinada etiqueta, com base em $P(t^k|t^j) = \frac{C(t^j, t^k)}{C(t^j)}$. Com o cálculo da probabilidade de $P(t_{i+1}|t_i)$, é possível determinar a probabilidade de uma seqüencia particular de etiquetas, ou melhor, a seqüencia de etiquetas de maior probabilidade para a seqüência de palavras considerada, ou ainda, o estado (no modelo de Markov, pois no domínio em questão, os estados do modelo de Markov são instanciados como etiquetas) de maior probabilidade para a seqüencia de palavras. As palavras são incorporadas no processamento com a transição de cada estado na cadeia de Markov. Segundo Jackson e Moulinier (2002), este comportamento assemelha-se à probabilidade de emissão de símbolos b_{ijk}, considerando que $P(O_n = k|X_n = s_i, X_{n+1} = s_j) = b_{ijk}$ e que cada etiqueta corresponde a um estado diferente.

É também possível estimar a probabilidade de uma palavra ser emitida por um estado (etiqueta) particular via a estimativa de máxima verossimilhança: $P(w^l|t^j) = \frac{C(w^l, t^j)}{C(t^j)}$. Com todos os elementos reunidos é possível determinar a melhor etiqueta $t_{1,n}$ para uma sentença $w_{1,n}$, através da aplicação do teorema de Bayes, que permite o cálculo de probabilidade condicionada (Jackson e Moulinier, 2002), conforme Equação 5.

$$arg_{t_{1,n}} max P(t_{1,n}|w_{1,n}) \quad \frac{= arg_{t_{1,n}} \max\left(P(w_{1,n}|t_{1,n})P(t_{1,n})\right)}{P(w_{1,n})} = \quad = arg_{t_{1,n}} max P(w_{1,n}|t_{1,n})P(t_{1,n})P(t_{1,n})$$

Equação 5

É possível reduzir a Equação 5, a partir de parâmetros estimados com base *corpus* de treinamento, adicionalmente, com base na propriedade de horizonte limitado, é possível assumir que (Manning e Schuetze, 1999):

a) palavras são independentes entre si (Equação 6), e

b) a identidade de uma palavra depende unicamente de sua etiqueta (Equação 7).

$$P(w_{1,n}|t_{1,n})P(t_{1,n}) \quad = \prod_{i=1}^{n} P(w_i|t_{1,n}) \mathrm{x} P(t_n|t_{1,n-1}) \mathrm{x} (t_{n-1}|t_{1,n-2}) \mathrm{x} \dots \mathrm{x} P(t_2|t_1)$$

Equação 6

$$P(w_{1,n}|t_{1,n})P(t_{1,n}) \quad = \prod_{i=1}^{n} P(w_i|t_i) \quad \mathrm{x} P(t_n|t_{n-1}) \mathrm{x} (t_{n-1}|t_{n-2}) \mathrm{x} \dots \mathrm{x} P(t_2|t_1) =$$

$$= \prod_{i=1}^{n} P(w_i|t_i) \mathrm{x} P(t_i|t_{i-1})$$

Equação 7

Como hipótese simplificadora Jackson e Moulinier (2002) apontam que $P(t_1|t_0)$ pode ser igualado a 1.0, gerando, por fim, a Equação **8**, que representa a equação final para a determinação a etiqueta ótima para uma sentença.

$$\hat{t}_{1,n} = arg_{t_{1,n}} max P(t_{1,n}|w_{1,n}) = arg_{t_{1,n}} max \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

Equação 8

Manning e Schuetze (1999) apontam o algoritmo Viterbi como uma eficiente alternativa para a etiquetagem de textos. O algoritmo em questão é utilizado como base em inúmeras soluções de PLN, como as ferramentas disponíveis sobre o guarda-chuva da OpenNLP (OPENNLP, 2009).

O algoritmo Viterbi é dividido em três etapas: a da inicialização, de indução e a de conclusão. Duas são as funções principais utilizadas, $\delta_i(j)$, que determina a probabilidade do estado inicial $j = (etiqueta\ j)$ da palavra *i*, e $\psi_{i+1}(j)$, que detemina o estado (ou etiqueta) mais provável da palavra *i*, com base na premissa que a máquina de estados está no estado *j* da palavra *i*+1.

Na etapa de inicialização (Figura 5) é atribuído a probabilidade 1,0 para o PERIODO, sendo que a $\delta_0(PERIODO) = 1{,}0$ e $\delta_0(t) = 0{,}0\ para\ t \neq PERIODO$. Assumindo que uma sentença é delimitada por períodos (identificado através das sentenças que compõem o texto).

1 **Comentário:** Entrada: sentença de tamanho *n*
2 **Comentário:** Inicialização
3 $\delta_0(PERIODO) = 1{,}0$
4 $\delta_0(t) = 0.0\ para\ t \neq PERIODO$
5 **Comentário:** Indução
6 **for** *i*:=0 **to** *n* **step** 1 **do**
7 **for** todas as etiquetas t^j **do**
8 $\delta_{i+1}(t^j) := max_{1 \leq k \leq T}[\delta_i(t^k)\ \mathbf{x}\ P(w_{i+1}|t^j)\ \mathbf{x}\ \mathbf{P(t^j|t^k)}]$
9 $\psi_{i+1}(j) := argmax_{1 \leq k \leq T}[\delta_i(t^k)\ \mathbf{x}\ P(w_{i+1}|t^j)\ \mathbf{x}\ \mathbf{P(t^j|t^k)}]$
10 **End**
11 **End**
12 **Comentário:** Conclusão
13 $X_n = argmax_{1 \leq k \leq T}\delta_n(j)$
14 **for** *i*:=0 **to** *n* **step** -1 **do**
15 $X_j = \psi_{i+1}(X_j + 1)$
16 **end**
17 $P(X_1, \ldots, X_n) = max_{1 \leq k \leq T}\delta_{n+1}(t^j)$

Figura 5 — Algoritmo de Viterbi para a etiquetagem de palavras

A etapa de indução (Figura 5) é baseada na **Equação 8**, sendo que $a_{jk} = P(t^k|t^j)$ e $b_{jkw^l} = P(w^l|t^j)$, conduzindo a $\delta_{i+1}(t^j) = max_{1 \leq k \leq T}[\delta_i(t^k)\ \mathbf{x}\ P(w_{i+1}|t^j)\mathbf{x}\ P(t^j|t^k)], \mathbf{1 \leq j \leq T}$ e $\psi_{i+1}(t^j) = argmax_{1 \leq k \leq T}[\delta_i(t^k)\ \mathbf{x}\ P(w_{i+1}|t^j)\ \mathbf{x}\ P(t^j|t^k)], \mathbf{1 \leq j \leq T}$.

Para a etapa de conclusão (Figura 5), assumindo $X_1, \ldots, X_n$ como o conjunto de etiquetas selecionado para o conjunto de palavras $w_1, \ldots, w_n$, calcula-se $X_n = argmax_{1 \leq j \leq T}\delta_n(t^j)$, $X_i = \psi_{i+1}(X_{i+1}), 1 \leq i \leq n-1$ e $P(X_i, \ldots, X_n) = max_{1 \leq j \leq T}\delta_{n+1}(t^j)$.

Jackson e Moulinier (2002), destacam a necessidade da adoção de estratégias específicas para o caso de palavras que não façam parte do modelo de treinamento e, por conseguinte, não tenham modelos conhecidos de probabilidade para o conjunto possível de etiquetas. Neste sentido, Weishedel, Meteer e Schuwartz (1993) apresentam estratégias específicas dependentes de idioma, no caso o inglês. Franz (1997) apresenta modelo de interdependência, com base em uma estratégia *loglinear*, focado na incorporação de palavras conhecidas e nas que são capitalizadas, em uma estratégia baseada em inferências Bayesianas, normalmente incorporado ao algoritmo de Viterbi (Figura 5) como no caso das ferramentas da (OPENNLP, 2009).

2.5 MorphoSaurus

A palavra, na Teoria Sintática (Paladino, 2006), é o elemento atômico na representação lingüística e, por conseguinte, forma a representação final dos objetos do mundo através da linguagem.

Nas sublinguagens, no entanto, é comum a presença de composições e outras derivações lingüísticas, como as expressões de domínio, evidenciando que a atomicidade semântica freqüentemente não coincide com o nível da palavra. Nas sublinguagens médicas, em especial, os sentidos atômicos são encontrados em diferentes níveis de fragmentação e granularidade. Um sentido atômico pode ser identificado através de um:

a) radical (ex. “hepat” refere-se a “fígado”);
b) prefixo (ex. “anti-”, ”hipo-”, ”des”);
c) sufixo (ex. “ose”, ”ite”, ”logia”);
d) fragmento (ex. “neurosis”, “hipofis-”); e
e) a palavra em si, seja ela composta ou não (ex. “pé”).

Termos compostos – tanto multi-palavra (ex. “vitamina A”) quanto aglutinadas (ex.; “hipofis-”), apresentam, normalmente, um sentido atômico que não pode ser derivado do significado dos componentes. Há também casos em que o significado do termo composto é totalmente diferente dos componentes, por exemplo, “neurose” (“neur” refere-se a nervos e “ose” a doença) não é uma doença de nervos e “*mycosis fungoides*” não é uma micose.

Considerando que as bases de documentos médicos, sejam de origem clínica ou literária, são normalmente compostas de diversas sublinguagens (ex.: a da enfermagem, a da fisioterapia, dentre outras), bem como a dimensão das bases disponíveis (apenas no Brasil mais de 360 milhões de consultas médicas são realizadas por ano (Carneiro, Leão e Pereira, 2002)) e a dinamicidade das mesmas (pois são constantemente atualizadas), a aplicação de muitas das abordagens tradicionais de Recuperação de Informação é dificultada e mostra resultados inferiores quando comparado com cenários concebidos sob condições experimentais de pequena escala como a indexação semântica latente ou modelos probabilísticos ainda mais sofisticados (Fuhr, 1992).

No caso das bases literárias, a recuperação de informação exige um conhecimento lingüístico e de domínio bastante apurado, dificultado quando da pesquisa em bases de documentos disponíveis em idioma que não o nativo do usuário. Neste cenário, como linha de ação para a adoção de técnicas de Recuperação de Informação no domínio médico, identifica-se a necessidade da inter-relação entre as diferentes terminologias, bem como o mapeamento

conceitual e, num nível mais detalhado, o mapeamento lexical de forma intra e inter-lingual, como apresentado em Pacheco, *et al*., 2007.

É neste domínio que o sistema MorphoSaurus foi construído, objetivando propor soluções computacionais para o multi-lingüismo e a granularidade semântica. Para tanto, são empregados descritores artificiais representativos dos conceitos da terminologia médica, identificados com base em num tesauro multilíngüe construído de unidades lexicais semanticamente atômicas. Em primeira instância, o sistema MorphoSaurus (Daumke, Schulz, *et al.*, 2009) pode ser entendido como uma ferramenta de normalização morfológica específica para a área médica, baseada em *subword indexing*.

Por *subwords*, entende-se elementos, fragmentos de palavras, que representam uma unidade lexical mínima no domínio (correspondendo a morfemas ou grupos de morfemas).

A metodologia principal do sistema MorphoSaurus consiste na submissão dos documentos a um processo de normalização morfossemântica (Schulz e Han, 2000) antes do processo de indexação, objetivando a melhora do desempenho no processo de recuperação.

Para o alcance deste objetivo, a normalização semântica faz uso de um tesauro médico. O tesauro é um conjunto de termos, em um determinado domínio, co-relacionados entre si, compostos por sinônimos e relações semânticas (Marques, 2002), ou seja, um vocabulário controlado da terminologia em um determinado domínio. O processamento baseado no tesauro melhora o processo de indexação e recuperação no domínio (Pacheco, Nohama, *et al.*, 2006).

Na construção do tesauro, os termos controlados são agrupados segundo seus sentidos atômicos em unidades lexicais. Os nós identificados são denominados de classe de equivalência ou ***MorphoSaurus ID****entifiers* - MID, sendo que as classes são construções independentes de idioma, ou seja, cada nó apresenta o mesmo significado em qualquer um dos idiomas considerados.

Para a definição de um MID, é necessário observar que uma seqüência de caracteres é considerada semanticamente atômica quando o seu significado não deriva unicamente de seus morfemas constituintes, seja por inflexão, derivação ou composição na formação de uma palavra; ou seja, ela por si só é representativa de um dado conceito (Pacheco, Nohama, *et al.*, 2006). Na verdade, é comum identificar palavras cujo significado é bastante diferente de seus morfemas originantes, por exemplo, enquanto **ose** significa doença, e **neuro**, nervo, a composição **neurose** indica uma doença de alcunha psicológica e não “doença dos nervos”.

De forma a representar conceitos atômicos, um MID baseia-se na construção de *subwords*, considerando cenários envolvendo **homonímia**, múltiplos sentidos, e a **sinonímia**,

mesmo sentido, múltiplas formas. Uma *subword* pode ser classificada como (Pacheco, Cancian, *et al.*, 2007)[10]:

a) *stem* - ST: aproximação lexical para radical[11], mas sem o formalismo do mesmo, exemplo: “gastr”, “hepat”, “diaphys”, “neuros”;
b) prefixo - PF: “afixo que vem antes da raiz (semantema); por ex., em **previdente** temos os morfemas **pre**- + **vid**- + -**ente**, onde **vid**- é o *stem* e **pre**- é o prefixo; em **imprevidente** temos dois prefixos, **im**- e **pre**-”;
c) prefixos próprios – PP: como “peri-”, “hemi-”, que representam prefixos que não podem ser prefixados;
d) sufixo - SF: “afixo que, posposto a uma raiz, radical, tema ou palavra, produz formas flexionadas ou derivadas”;
e) sufixos próprios - SFP: como “-ação”, “-ão”, “-essemos”, são sufixos que não podem ser acompanhados de outros sufixos;
f) invariante - IV: como “gastr-o-intestinal”, ou “- r -” em “hernio-r-rafia” são usados (motivação fonológica) como um ente de ligação entres *stems*;

As *subwords* como “-logia” e “-itis”, com pesos semânticos consideráveis, classificados como sufixos são controversas. Como regra geral, o critério para se assumir como *stem* é que este não precise de outro *stem* para a boa formação de uma palavra ou expressão. Por questões de funcionalidade do sistema, são classificados como tal. E, por fim, invariantes como “ion”, “gen”, ou nomes próprios como “aspirina” e acrônimos como ECG ou AIDS; coincide com as palavras que não permitem segmentação. Em muitos casos, coincide com palavras muito curtas que podem apresentar problemas de ambigüidades se forem utilizadas como unidades atômicas para construção de termos complexos.

As bases terminológicas que compõem o sistema MorphoSaurus são constituídas pelos seguintes componentes (Schulz e Hahn, 2006):

a) um repositório de *subwords*, no qual a delimitação de *subwords* se efetua de tal modo que sirvam para a resolução de sinônimos, prevenindo, ao mesmo tempo a proliferação de ambigüidades no processo de segmentação do texto;
b) um repositório de nomes próprios, tais como nomes de medicamentos ou nomes próprios, utilizados dentro de termos médicos, por exemplo, doença de *Alzheimer*; e

[10] A classificação de entidades morfológicas não corresponde a uma tipologia lingüística, mas foi desenvolvida em dependência da implementação concreta da segmentação, para a qual a extração de sensos das palavras é o único objetivo o que justifica qualquer desvio dos princípios lingüísticos.
[11] Relativo ou pertencente à raiz ou à origem (Ferreira, 1999).

c) um componente de tesauro que agrupa *subwords* de significado idêntico assinalando-lhes um identificador comum.

O processo de indexação, baseado no processamento de MIDs, é dividido em:

a) normalização ortográfica (Figura 6), que consiste na substituição de caracteres especiais para cada língua (exemplo: “ß” → “ss”, no alemão e “á” → “a”, no português) e na normalização de palavras capitalizadas, bem como outros processamentos específicos para cada idioma;

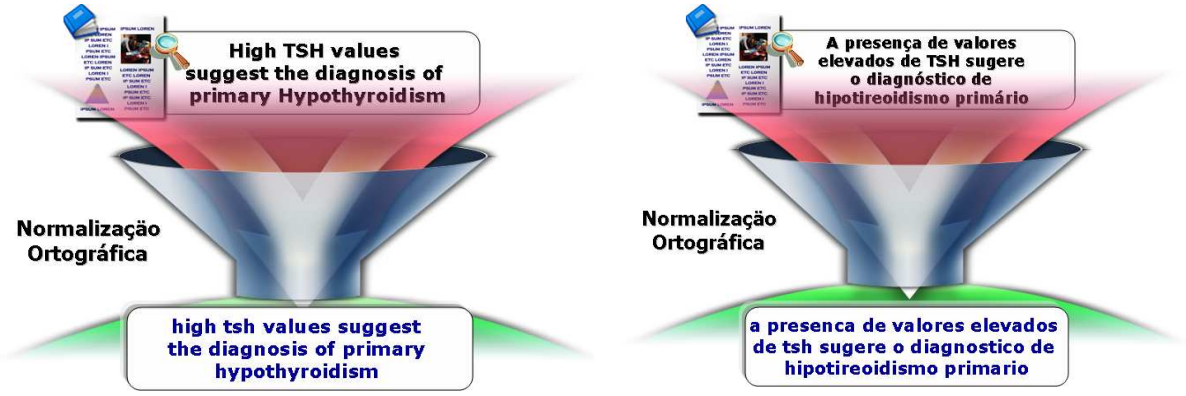


Figura 6 — Normalização Ortográfica aplicado a dois exemplos comparáveis

b) na segmentação morfológica (Figura 7), o texto normalizado é decomposto em uma seqüência de *subwords*, pela ação de um autômato finito, através da aplicação de mecanismos de *hash*, tendo como base o léxico. A função de *hash* prioriza o *match* da seqüência mais longa (caso não seja identificado um elemento no léxico, a palavra original é retornada);

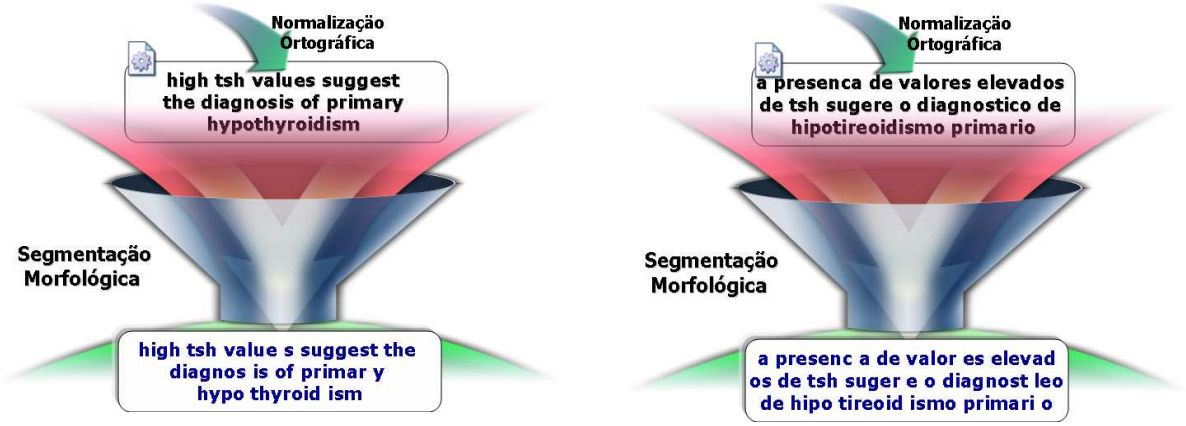


Figura 7 — Segmentação Morfológica aplicado a dois exemplos comparáveis

c) na normalização semântica (Figura 8), cada *subword* é substituída pelo respectivo MID. Desta forma, todas as traduções de um termo, em diferentes línguas, são representadas pelo mesmo MID, bem como todas as sinonímias de um idioma específico.

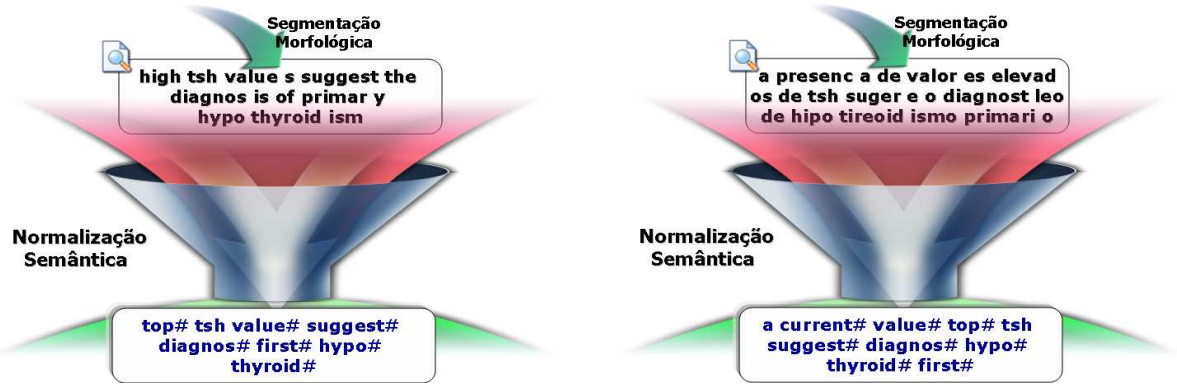


Figura 8 — Normalização Semântica aplicada a dois exemplos comparáveis (inglês e português)

Procurando formalizar a definição de uma *subword* estabelece-se um conjunto de elementos lexicais atômicos, ***LU***, fixa-se ***M*** como sendo um conjunto de classes de equivalência, gerado a partir de uma função injetora ***f,*** e não ambíguo de MID, convencionalmente precedido do símbolo #. Cada entrada lexical é denotada com um tipo ***T*** := {PP, PF, ST, IV, SF, SP}, sendo ***L*** a representação dos idiomas contemplados no tesauro, que atualmente contempla o Inglês, Alemão, Francês, Espanhol, Português e Sueco, e, por fim, um identificador de domínio ***D***, completando o quíntuplo formado (***LU, M, T, D, L***) (Andrade, 2007; Schulz e Hahn, 2006), sendo **M = f(LU)**.

No caso da não existência de significado relacionada a uma entrada lexical, a mesma é tratada como "*stop entry*", sendo desconsiderada no processamento. Caso típico para elementos com papel somente gramatical, como terminações e verbos auxiliares, como também para termos não cobertos pelo léxico.

Se nenhum significado for assinalado para a entrada lexical, então, esta é considerada como uma "*stop entry*" (termo não considerado no processamento); tendo somente uma função gramatical, como, por exemplo, os verbos auxiliares e terminações utilizadas nas inflexões das palavras. Alguns exemplos típicos de entradas lexicais e suas inter-relações correspondentes no MorphoSaurus (Andrade, 2007; Pacheco, Nohama, *et al.*, 2006):

a) Para o caso de sinonímia, o sufixo em inglês “-itic” e ”-itis” possuem o mesmo sentido de “*inflammation*”:

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

b) Para o cenário de tradução, é possível destacar o caso do *stem* alemão *entzünd* (transcrito para *entzuend*) e o sufixo em português “-ite” denota o mesmo sentido do *stem* inglês “*inflamm*”:

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

c) Nos casos ambíguos, destaca-se o substantivo “*head*” em inglês, ou “cabeça” em português, pode se referir a uma parte anatômica do corpo como a uma pessoa:

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

d) A palavra “*era*” é um substantivo em inglês, mas um verbo auxiliar nas línguas latinas, denotando o caso de um “s*top entry”*

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

e) Existem casos de quase-sinônimia, em que um termo é equivalente a outro apenas em algum contexto específico, como as palavras “sildenafil” e o nome “viagra” podem ser considerados sinônimos para medicina clínica (d_1), mas não no campo da indústria farmacêutica (d_2):

   i. L_{11} = (*sildenafil*, ST, #*sildenafil* , EN, d_1)
   ii. L_{12} = (*viagra*, IV, #*sildenafil* , EN, d_1)
   iii. L_{13} = (*sildenafil*, ST, #*sildenafil* , EN, d_2)
   iv. L_{14} = (*viagra*, IV, #*viagra*, EN, d_2).

2.5.1. Construção do tesauro do MorphoSaurus – história e características

O sistema MorphoSaurus é um projeto internacional que envolve os grupos de informática médica da PUCPR e da UTFPR, no Brasil, e das Universidades de Freiburg e Jena, na Alemanha.

O tesauro foi construído com base em *corpus* de domínio. *Corpus* é "uma coletânea de textos escritos em linguagem natural (*naturally occuring*), escolhidos para caracterizar um estado ou variedade da linguagem" (Sinclair, 1995).

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

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

Quanto à **extensão do *corpus***, o mesmo deve ser uma coletânea "representativa e criteriosa de textos naturais" (Percy e Lancashire, 1996). Por criteriosa, entende-se que deva refletir a variedade o mais fielmente possível; ou seja, para um *corpus* geral de uma língua, deve-se incluir a maior quantidade de ocorrência de palavras possíveis no domínio em questão.

A linguagem é um sistema probabilístico (Halliday, 2001) no qual certos traços são mais freqüentes que outros. No caso do léxico podem-se diferenciar as palavras entre aquelas de maior freqüência e as de menor freqüência, sendo que a diferença entre elas é relativa. Assim, algumas palavras têm freqüência de ocorrência muito rara e, para que haja probabilidade de ocorrência no *corpus*, é necessário incorporar uma grande quantidade de palavras. Portanto, quanto maior a quantidade de palavras, maior a probabilidade de ocorrência de palavras com baixa freqüência.

No caso dos sentidos das palavras, também é possível distinguir entre os sentidos mais freqüentes e os menos freqüentes dos itens lexicais. Assim, mesmo palavras de alta freqüência têm sentidos raros (por exemplo, “cabeça” entendida como gestor de uma organização) que terão maior probabilidade de ocorrer quanto maior for o *corpus*.

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

A freqüência em si não é suficiente, porque mesmo palavras de alta freqüência possuem vários sentidos. Assim, uma freqüência alta pode esconder vários sentidos, que separados teriam baixa freqüência. Para que seja representativo, um *corpus* deve conter o maior número possível de sentidos de cada forma. Por exemplo, a forma “como” pode significar a preposição ou a primeira pessoa do singular do verbo comer no presente do indicativo.

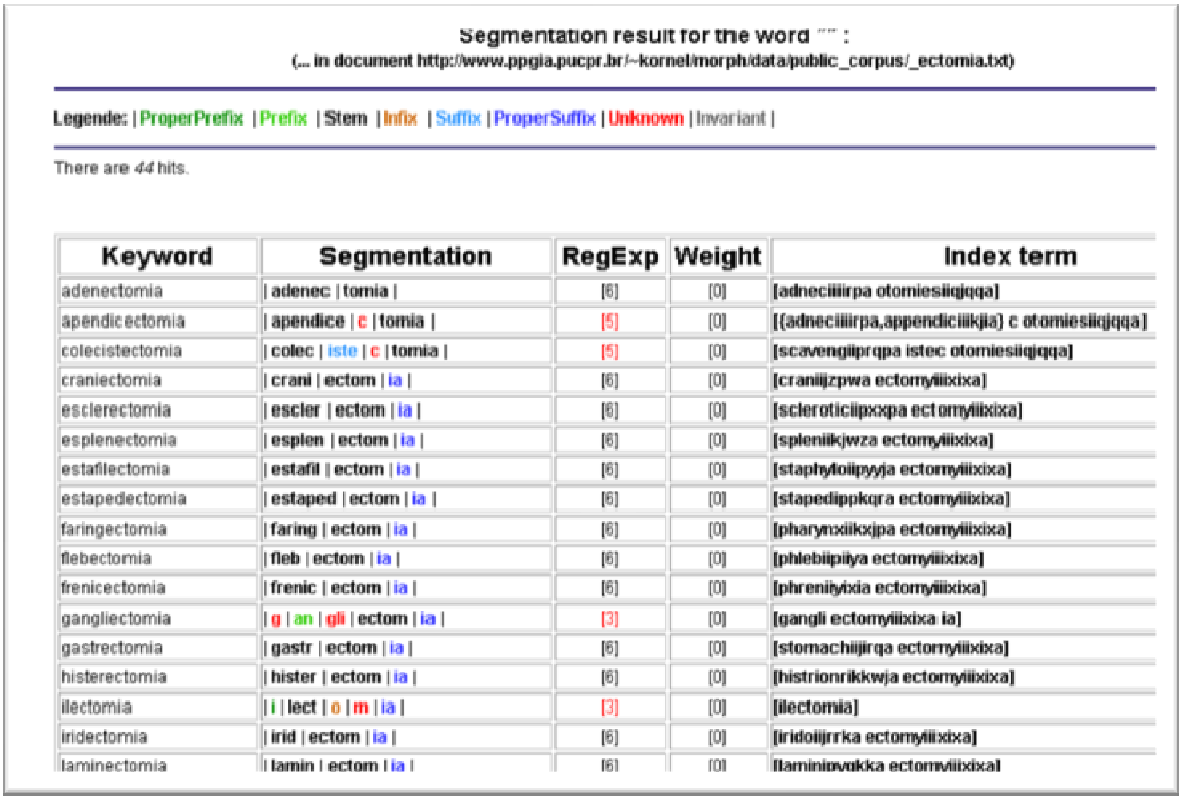
Para a criação, a agregação de termos no domínio do thesaurus do MorphoSaurus, foi produzida uma lista de termos versus freqüência do *corpus*, com base em diferentes domínios e linguagens. Aqueles com índice de freqüência muito alta normalmente se comportavam como *stopwords* e foram eliminadas. As de baixas freqüências normalmente eram motivo de melhores investigações por se tratar em alguns casos de termos específicos da área de análise.

Ao longo dos ciclos de trabalho, a segmentação era rotineiramente verificada através do segmentador (Figura 9 e Figura 10).

A edição do tesauro era, inicialmente, realizada através de um aplicativo local, com base de dados MSAccess. A incorporação de novos idiomas (atualmente: alemão, inglês, português, espanhol, francês, sueco e italiano), bem como a inter-relação de outras equipes (Suécia e Itália), demandou o desenvolvimento de uma aplicação Web, conhecida como MorphoEditWeb[12] (Figura 11).

[12] http://morphwww.medinf.uni-freiburg.de/

Figura 9 — Interface do "MorphoSaurus *Segmenter*" utilizado para simulações de segmentação de palavras



Keyword	Segmentation	RegExp	Weight	Index term
adenectomia	| adenec | tomia |	[6]	[0]	[adneciiiirpa otomiesiiqjqqa]
apendicectomia	| apendice | c | tomia |	[5]	[0]	[(adneciiiirpa,appendiciiikjia) c otomiesiiqjqqa]
colecistectomia	| colec | iste | c | tomia |	[5]	[0]	[scavengiiprqpa istec otomiesiiqjqqa]
craniectomia	| crani | ectom | ia |	[6]	[0]	[craniijzpwa ectomyiiixixa]
esclerectomia	| escler | ectom | ia |	[6]	[0]	[scleroticiipxxpa ectomyiiixixa]
esplenectomia	| esplen | ectom | ia |	[6]	[0]	[spleniikjwza ectomyiiixixa]
estafilectomia	| estafil | ectom | ia |	[6]	[0]	[staphyloiipyyja ectomyiiixixa]
estapedectomia	| estaped | ectom | ia |	[6]	[0]	[stapediippkqra ectomyiiixixa]
faringectomia	| faring | ectom | ia |	[6]	[0]	[pharynxiiixxjpa ectomyiiixixa]
flebectomia	| fleb | ectom | ia |	[6]	[0]	[phlebiipiiya ectomyiiixixa]
frenicectomia	| frenic | ectom | ia |	[6]	[0]	[phreniiyixia ectomyiiixixa]
gangliectomia	| g | an | gli | ectom | ia |	[3]	[0]	[gangli ectomyiiixixa ia]
gastrectomia	| gastr | ectom | ia |	[6]	[0]	[stomachiijirqa ectomyiiixixa]
histerectomia	| hister | ectom | ia |	[6]	[0]	[histrionrikkwja ectomyiiixixa]
ilectomia	| i | lect | o | m | ia |	[3]	[0]	[ilectomia]
iridectomia	| irid | ectom | ia |	[6]	[0]	[iridoiijrrka ectomyiiixixa]
laminectomia	| lamin | ectom | ia |	[6]	[0]	[laminiovxkka ectomyiiixixa]

Figura 10 — Exemplo de saída do segmentador aplicado em um arquivo de entrada

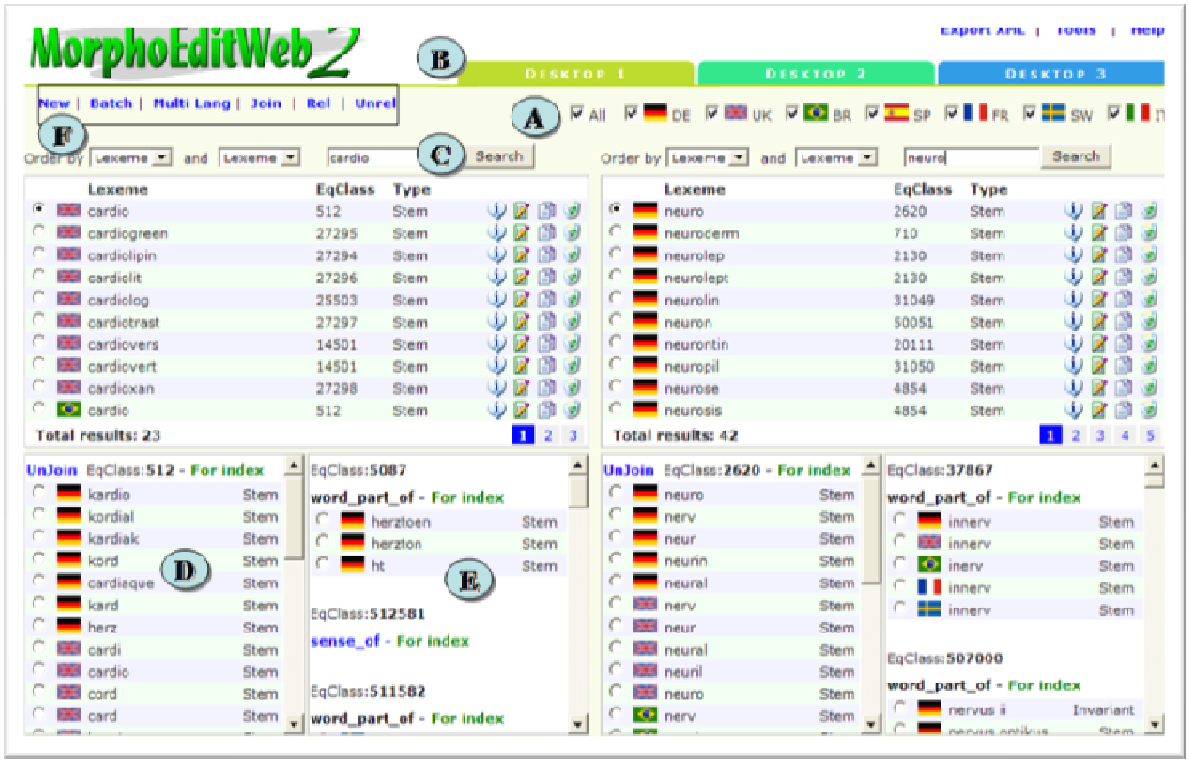


Figura 11 — Interface gráfica do MorphoEditWeb[13]

2.6 Inteligência Artificial

Segundo Thro (1991), a IA é uma ciência "multi e interdisciplinar que aborda a neurociência, filosofia, psicologia, matemática, estatística, ciência da computação, robótica e lingüística computacional; e se dedica à reprodução dos métodos ou resultados do raciocínio humano e da atividade do cérebro". Todavia, esta definição apresenta significados diferentes para pessoas diferentes. As expectativas da IA na neurociência ou na psicologia são muito diferentes das expectativas da IA na ciência da computação ou na robótica (Waterman, 1986).

Desde que o termo "inteligência artificial" foi apresentado por Minsky na conferência de Dartmouth em 1956, várias definições surgiram para delimitar o campo de atuação desta nova ciência (Waltz, 2005). No contexto do presente trabalho, o domínio de interesse envolve apenas o campo da Ciência da Computação e, neste contexto, a IA visa o entendimento e a implementação da capacidade de simular características da inteligência humana, objetivando a

[13] As principais funcionalidades do MorphoEditWeb (Figura 11) são: A) manipulação de lexemas em múltiplos desktops; B) filtros baseados em linguagem; C) filtros baseados em palavras chave; D) exibição da composição de uma Classe de Equivalência; E) edição das relações entre classes de equivalência, e F) ações de manipulação de lexemas, como: criação de nova entrada na base, união entre classes de equivalência existentes, entre outras.

solução de problemas que não apresentem solução satisfatória pelos métodos convencionais (Waltz, 2005).

2.6.1. Agentes

Uma variedade de definições de agentes tem sido apresentada por investigadores da área da IAD, cada qual orientada aos seus propósitos.

Russel e Norvig (1995) definem que:

> "Um agente é uma entidade que pode perceber o seu ambiente através de sensores e agir sobre este ambiente através de atuadores".

Maes (1995) destaca que:

> "Agentes autônomos são sistemas computacionais, os quais inseridos num ambiente dinâmico e complexo, percebem e atuam automaticamente neste ambiente, e fazendo-o, compreendem um conjunto de objetivos ou tarefas para as quais foram projetados".

Em Smith, Cypher e Sopher (1994) é declarado:

> "Definimos agente como uma entidade persistente de software dedicada a uma finalidade específica. O termo persistente é empregado para distinguir agentes de rotinas; agentes possuem suas próprias idéias de como devem realizar suas tarefas e as suas próprias agendas".

Hayes (1995) descreve:

> “Agentes inteligentes executam de forma contínua três funções: (i) perceber as condições dinâmicas do ambiente, (ii) agir para afetar as condições do ambiente e, (iii) raciocinar para interpretar as percepções, resolver problemas, inferir e determinar ações”.

Para Wooldridge e Jennings (1995):

> “Agente é utilizado para denotar um sistema computacional que goza das seguintes propriedades: (i) autonomia: agentes operam sem a intervenção direta de seres humanos ou outras entidades, e exercem algum tipo de controle sobre as suas ações e estados internos, (ii) habilidade social: agentes interagem com outros agentes, (iii) reatividade: agentes percebem os seus ambientes e respondem rapidamente às trocas que neles ocorrem e, (iv) pró-atividade: agentes não agem apenas em resposta às alterações dos seus ambientes. Eles

> são também capazes de exibir um comportamento orientado por objetivos através de iniciativas".

Adelinde e Weyns (2009) definem:

> "Um agente é definido como um sistema computacional que é capaz de ações autônomas no ambiente conforme sua especificação original. Autonomia, neste cenário, identifica que o sistema deva ser capaz de operar sem intervenção direta (seja humana ou de outros agentes), tendo controle sob seu estado interno".

Em síntese, agentes são programas de computador habilitados para responder às solicitações específicas sobre o domínio de conhecimento do mesmo; sendo que a observação dos fenômenos do mundo é realizada através de um conjunto de sensores e a atuação ocorre através do auxílio de um conjunto de atuadores.

A analogia feita com agentes no mundo real permite conceituar um agente como uma entidade ativa que possui conhecimento especifico sobre um determinado domínio.

Wooldridge (Wooldridge, 1995) apresenta um conjunto de propriedades desejáveis a um agente, a saber:

a) **autonomia**: o Agente Inteligente deve apresentar a capacidade para determinar as ações principais para conclusão de tarefas ou objetivos, sem a intervenção manual, representando um fator de independência para o Agente;

b) **habilidade de comunicação**: Agentes Inteligentes costumam acessar informações da fonte de terceiros sobre o estado corrente do mundo externo. Isto requer uma habilidade para comunicar com as bases de dados desta informação. Esta comunicação pode ser na forma de uma inspeção singular com um simples e conciso grupo de possibilidades de respostas ou uma comunicação complexa com várias respostas. A comunicação, entretanto, pode ocorrer em um nível mais alto, envolvendo verdadeiros diálogos. Agentes revelam intenções e objetivos, eventualmente baseados em algum acordo ou "contrato";

c) **capacidade para cooperação**: a extensão natural do atributo Comunicação é a Cooperação. Agentes Inteligentes devem ter uma participação colaboradora para existir e suceder em um Sistema Orientado a Agente. A meta é que os Agentes trabalhem juntos para a realização de atividades;

d) **capacidade de raciocínio**: a habilidade para raciocinar é um dos aspectos chaves de um Agente Inteligente. Raciocinar implica que um Agente deva possuir a habilidade para deduzir e extrapolar baseado em conhecimento corrente e experiência; e

e) **procedimento adaptado**: já que os Agentes Inteligentes devem ter a capacidade de serem autônomos e demonstrar razão, eles devem apresentar mecanismos para serem capazes de acessar os estados correntes de seu domínio externo, que pode ser definido como a extensão do conhecimento ao alcance do Agente e incorporá-lo entre suas decisões sobre ações futuras. Agentes devem demonstrar habilidade para examinar o conhecimento externo e o sucesso de previsão de ações tomadas de condições similares, adaptando suas ações para aperfeiçoamento da probabilidade de sucesso.

A partir das características enunciadas, conclui-se que a limitação dos recursos computacionais disponíveis e seu tipo particular de concepção podem dar a um Agente certo conjunto particular de características. Portanto, é importante que se defina o modelo que se busca para a construção de um Agente, bem como a arquitetura objetivando o exercício de controle sobre os recursos disponibilizados, permitindo a integração das características apresentadas.

2.6.2. Sistemas Multiagente

De forma geral, o termo sistemas multiagente tem sido aplicado a qualquer sistema composto de múltiplos agentes interatuantes (Shmeil, 1999; Wooldridge, 1995; Adelinde e Weyns, 2009). De forma específica, um sistema multiagente é composto de: um ambiente *Am*, um conjunto *Ag* de agentes e um conjunto *Ob* de objetos, não agentes.

Um ambiente *Am* é um espaço dinâmico, dotado de uma métrica que possibilita os agentes perceberem, localizarem e atuarem sobre os objetos. Um conjunto de agentes *Ag* é um conjunto que apresenta uma estrutura de organização, de agentes homogêneos ou heterogêneos, os quais agem/reagem no ambiente *Am* diante dos elementos do conjunto *Ob* ou face aos elementos do conjunto *Ag*, através de **comportamentos**. Um conjunto *Ob* de objetos é um conjunto de entidades, não agentes, presentes no ambiente *Am* que sofrem manipulações pelos elementos do conjunto *Ag*.

Um sistema multiagente tem uma estrutura de organização de **nível de sociedade** quando o foco recai sobre num grande número de agentes, seus múltiplos papéis, atividades e evolução na comunidade como um todo. É do **nível de grupo** quando o interesse está concentrado nas relações entre agentes, seus papéis e atividades, agregados em torno de um segmento da sociedade. É do **nível micro** quando a ênfase ocorre essencialmente nas relações entre dois ou entre um reduzido número de agentes, os quais representam um subconjunto de um grupo (Wooldridge, 1995; Adelinde e Weyns, 2009).

Considerando as definições de agentes apresentadas no presente trabalho, bem como as definições em geral, apresentam propriedades as quais, além de permitirem a distinção e caracterização de entidades, no caso agentes, em relação a outras entidades, possibilitam a formação de classes. Estas classes são baseadas em dimensões (concepção, constituição, granularidade, etc.), e são úteis para o enquadramento dos agentes, permitindo classificar o sistema multiagente em sistema **heterogêneo** ou **homogêneo** dependendo da dimensão em análise.

Um sistema multiagente é homogêneo quando os agentes participantes possuem, nas dimensões em análise, valores aceitáveis e congruentes. Um sistema multiagente é heterogêneo quando os agentes participantes possuem, nas dimensões em análise, valores aceitáveis e não congruentes. A Figura 12 apresenta um de entre vários esquemas possíveis (Huhns e Munindar, 1998) de classificação. Este esquema não é exaustivo, na medida em que apresenta apenas algumas das dimensões possíveis para uma completa taxonomia de agentes.

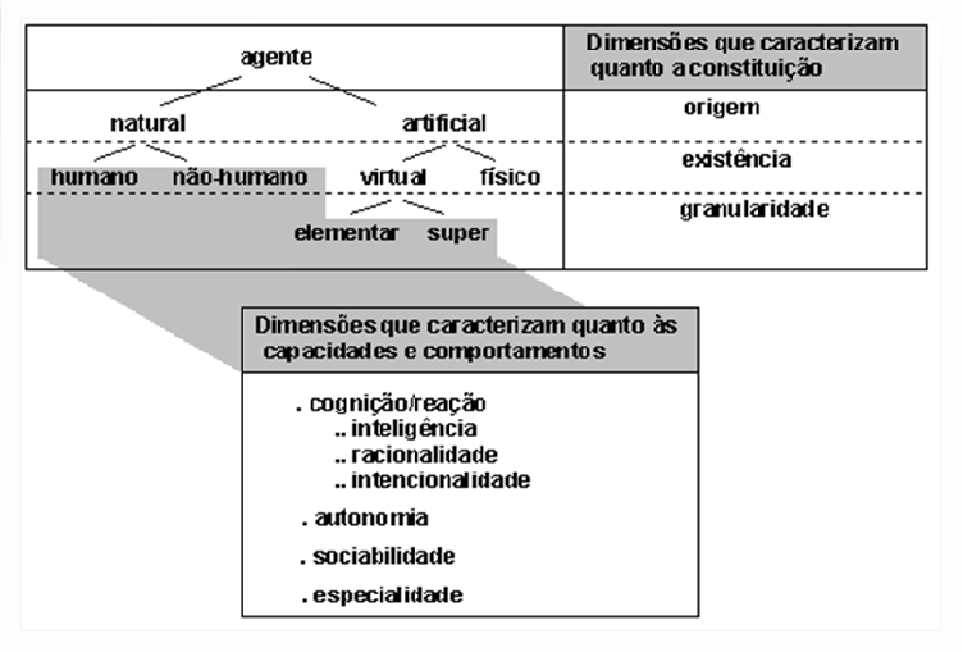


Figura 12 — Algumas das dimensões para classificação de agentes

Em referência à Figura 12, há as dimensões que caracterizam os agentes quanto à constituição e quanto às capacidades e comportamentos.

Quanto à constituição dos agentes (Shmeil, 1999):

a) **origem**: enquadra os agentes quanto à concepção, podendo ser: natural ou artificial. É natural quando o agente está contido nos objetos ou fenômenos estudados pelas ciências naturais. É artificial quando é um artefato "*man-made*" (Simon, 1968);

b) **existência**: os agentes artificiais, quanto à sua existência no mundo real, podem ser: física ou virtual. É física quando altera o mundo real (ex. robô, automóvel, etc.). É virtual quando se trata de um componente de software que representa um agente; e

c) **granularidade**: enquadra os agentes em termos de constituição quantitativa, podendo ser: elementar ou super. É elementar quando os componentes do agente não são qualificados como agentes. É super quando o agente é constituído por partes menores as quais são agentes elementares ou super agentes.

Quanto às capacidades e comportamentos dos agentes (Shmeil, 1999):

a) **cognitivos (ou deliberativos):** a hipótese de símbolos físicos de Simon é base para a maioria dos modelos de Agentes usados na IA e tem como núcleo a representação do ambiente e dos Estados Mentais. Estes estão sujeitos às alterações pelas inúmeras formas do raciocínio simbólico. As *crenças* de um Agente revelam o que ele espera do estado atual do ambiente e mostram suas expectativas sobre o curso das ações que deverão ser executadas para que ele possa atingir um determinado objetivo. As crenças são modeladas tendo como base a concepção de ambientes possíveis, associadas a cada tipo de representação, percebida por um Agente, de um determinado ambiente. De maneira abstrata, os *desejos* especificam preferências sobre estados futuros ou do curso dos acontecimentos do ambiente. Eles não são necessariamente consistentes e um Agente pode crer que um desejo venha a acontecer ou não. As formações de desejos através dos processos de seleção de objetivos são representadas pela *intenção*. Os cognitivos são chamados de intencionais na medida em que a estrutura de conhecimento para a aplicação do raciocínio se caracteriza pela existência de crenças, desejos e intenções

("*BDI-architecture, Beliefs, Desires and Intentions*"). Esta estrutura do conhecimento expressa uma atitude mental do agente (Wooldridge e Jennings, 1994), representando respectivamente a informação, a motivação e o raciocínio (referidos também na literatura como agentes deliberativos);

b) **reativos:** Fortemente baseados na psicologia comportamental, estes Agentes formam suas decisões em tempo de execução. Eles baseiam-se geralmente em pouca informação e através de regras de ação simples tendo como filosofia de trabalho a hipótese de Simon, que expressa que "a complexidade do comportamento de um Agente pode ser uma reflexão da complexidade do ambiente no qual este Agente esta operando mais do que seu modelo interno" (Simon, 1968). São reativos na medida em que percebem o ambiente no qual estão inseridos e respondem rapidamente às trocas ocorrem no meio. O termo rapidamente está associado à idéia de uma baixa atividade de raciocínio (permitindo reações rápidas) o que leva a caracterizá-los como incapazes de manipularem os seus objetivos, i.e. suas ações são executadas como o resultado de disparos de regras simples, dado pelo par (percepção, reação). Quanto à inteligência dos agentes reativos, esta emerge não do comportamento individual, mas do comportamento global da comunidade;

c) **sociabilidade**: enquadra os agentes em termos de possuírem explicitamente ou não, modelos dos outros agentes da comunidade, e através do raciocínio, considerarem este conhecimento para a tomada de decisão, e

d) **especialidade**: enquadra os agentes em termos do domínio das tarefas que os mesmos executam.

2.6.3. *Resource Description Framework - RDF*

A linguagem *eXtensible Markup Language* - XML (BRAY *et al.*, 2004) apresenta-se como alternativa para a integração de dados, mas nos seus limites não apresenta semântica. Por isso, tem-se assistido à construção de várias linguagens complementares que objetivam a resolução dessas lacunas. Neste domínio, o RDF, uma recomendação do W3C, constitui-se em uma arquitetura genérica de metadados que permite descrever recursos no contexto Web, através da adoção de padrões de metadados. "A idéia básica de RDF não é definir um conjunto universal de metadados e sim prover os mecanismos necessários para que as diversas

comunidades codifiquem, troquem e reutilizem metadados estruturados" (LASSILA e SWICK, 2006).

RDF pode ser representada através de "XML para a codificação de meta-informação, permitindo a descrição de recursos com o intuito de facilitar o processamento automático de informação, sem assumir nenhuma aplicação particular ou domínio semântico" (HAROLD e MEANS, 2000). Consistindo da descrição de nodos e dos seus pares atributo-valor. Um nodo é um identificador de recurso uniforme ou *Uniform Resource Identifier* - URI e representa um recurso qualquer. Os atributos são propriedades dos nodos e os seus valores são elementos de texto ou outros nodos.

O modelo RDF Schema (BROEKSTRA e KAMPMAN, 2005), que se baseia no modelo RDF básico, é fortemente influenciado por conceitos de orientação a objetos e de linguagens de especificação de bancos de dados.

As descrições que seguem encontram-se baseadas em dois documentos: *Resource Description Framework (RDF) Model and Syntax Specification* (LASSILA e SWICK, 2006), que descreve o modelo de dados RDF e *Resource Description Framework (RDF) Schema Specification*, que apresenta as primitivas de modelagem para a definição de um domínio particular de interesse.

2.6.4. *Unstructured Information Management Architecture*

Aplicações "*Unstructured Information Management*" - UIM são softwares que analisam informação não estruturada, como texto, áudio, vídeo, imagens, etc, objetivando a descoberta, organização e recuperação de conhecimento relevante. Para a consecução deste objetivo, aplicações UIM fazem uso de uma variedade de tecnologias de análise, incluindo PLN estatístico, Recuperação de Informação, *Machine Learning* e ontologia.

O "*Unstructured Information Management Architecture*" - UIMA é um *framework* que suporta a criação, descoberta,, composição e desenvolvimento de um conjunto de artefatos para análise e correlação com informação estruturada, como bancos de dados ou máquinas de busca. O UIMA foi criado e é mantido pela IBM (CREESE, 2005).

A tecnologia do UIMA é baseada na plataforma Java. O *framework* aplica técnicas de IA na sua arquitetura, em especial na área de agentes de software. Na Figura 13, apresenta-se a arquitetura do UIMA. Basicamente, o *framework* é baseado em blocos denominados de "Motor de Análise" (MA). O coração de um MA está nos algoritmos que analisam documentos e registram os resultados obtidos (por exemplo, detecção de nomes próprios).

Para permitir o reuso, UIMA define Adaptadores de Domínio, que é um agente de software genérico, que implementa um *container* baseado em OOP, possibilitando o gerenciamento e armazenamento de objetos, assim como suas propriedades e estados.

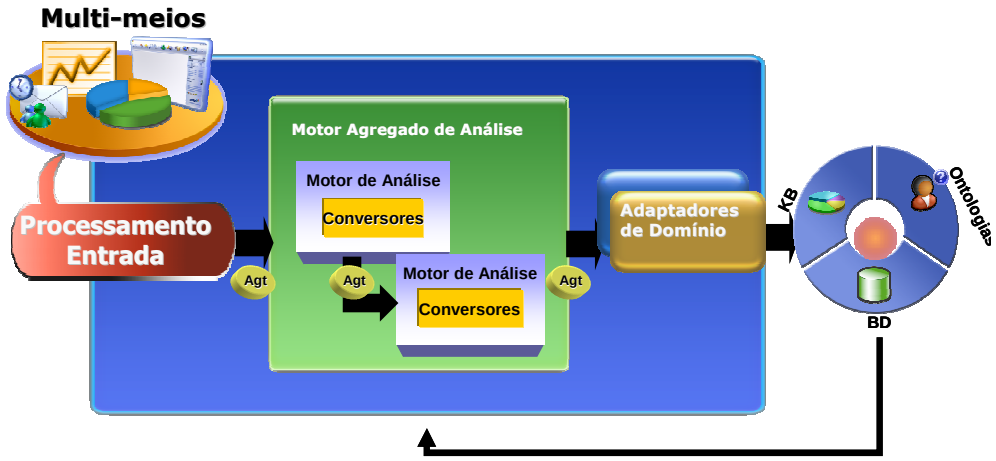


Figura 13 — Arquitetura do *framework* UIMA, ilustrando as diferentes etapas envolvidas no processamento de um documento na estrutura do *framework*

É possível encadear a execução de vários adaptadores de domínio, em uma comunidade multiagente. A comunicação e inter-relação entre os agentes são realizadas mediante a anotação das competências individuais e o fluxo de mensagem entre os agentes envolvidos.

2.7 Conclusões

A modelagem de um sistema com controle centralizado que considere a observação e atuação nos múltiplos e diferentes proponentes fenomenológicos, que formam o prontuário do paciente, conduziria a um conjunto de problemas relativo à elevada concentração de decisões a serem tomadas em um único componente, demandando capacidade computacional desproporcional e altamente custosa.

Neste cenário a modelagem do problema através de metodologias de Sistemas Multiagente, mostra-se adequada na medida em que:

a) aproxima-se do modo como na realidade este problema é formulado;

b) diminui a complexidade dos componentes, em razão da individualização dos componentes, transformando-os em unidades funcionais isoladas e intercomunicantes;

c) aumenta a flexibilidade de alterações, inclusões, eliminações e movimentações das unidades funcionais;
d) permite uma maior eficiência, usufruindo do partilhamento do tempo através do paralelismo;
e) distribui o tratamento de falhas, permitindo a adoção de estruturas alternativas para a solução de problemas, quando do acontecimento de falha; e
f) permite o balanceamento da carga computacional através da distribuição dos componentes de software.

Com relação à metodologia de representação de conhecimento, mais do que uma decisão técnico-estratégica, deve-se considerar a disponibilidade de soluções e tendências de mercado da área objeto de pesquisa. Neste sentido, as experiências científicas atuais apontam para o uso de ontologias. No caso específico do domínio médico, cabe ressaltar que a construção de uma base de conhecimento, independente da estratégia de representação é uma atividade multi-disciplinar e de grande monta.

Capítulo 3

3 METODOLOGIA

Neste capítulo, descrevem-se os procedimentos utilizados neste estudo para dar cumprimento aos objetivos propostos. No sentido de tornar mais clara a sua apresentação, dividiu-se o capítulo em sub-capítulos que contemplam: a descrição do estudo; a caracterização da amostra e os critérios que presidiram à sua seleção; o instrumento utilizado para coleta dos dados, com descrição dos processos seguidos na sua construção; as condições em que se efetuou a coleta de dados; e, por último, o tratamento dos dados.

3.1 Caracterização do domínio

O prontuário do paciente reúne um conjunto de informações sobre um ou mais episódios de atendimento clínico-hospitalar. Apesar de crescente e constante informatização das instituições de saúde, bem como do progresso que tem se identificado na área de terminologias e vocabulários controlados, documentos em texto livre ainda predominam, mesmo nos sistemas de informação mais avançados (Massad e Marin, 2003).

Essas narrativas são produzidas pelos próprios profissionais de saúde, os quais, por pressão de tempo e por dificuldade de grafia, freqüentemente acabam produzindo textos que só parcialmente correspondem às normas gramaticais e ortográficas. Isso é tolerado enquanto esses documentos servirem, predominantemente, à comunicação entre os profissionais e à documentação interna na instituição (Silva e Tavares-Neto, 2007). No entanto, tais características dificultam o processamento computacional desses documentos.

Utilizando tecnologias de PLN e gestão de conteúdo, esse processamento tem as finalidades mais variadas, incluindo recuperação de informação para identificação de

candidatos a estudos clínicos, ou casos similares, codificação automatizada de doenças e procedimentos, levantamentos epidemiológicos, apoio à decisão médica, assim como o controle e gestão de qualidade.

As tecnologias de PLN vigentes geralmente pressupõem textos que correspondem às normas lingüísticas. A quase totalidade das pesquisas em PLN baseia-se em textos jornalísticos, os quais exibem pouca incidência de erros. Ao contrário do leitor humano, os processos computacionais são mais susceptíveis a erros e pequenas interações que podem gerar resultados insatisfatórios (Sager e Lyman, 1994).

Objetivando explorar o processamento computacional de prontuários eletrônicos, optou-se pelo uso de sumários de alta. A escolha deste tipo de documento, em detrimento às evoluções ou laudos, deu-se pela abrangência e multidisciplinaridade deste tipo de documento, que objetiva a sumarização do conteúdo observado no prontuário do paciente, focando a comunicação médico-paciente (Klück e Guimarães, 1999).

As seções de um sumário de alta incluem, normalmente, exames físicos, antecedentes pessoais e familiares, sintomas e sinais observados na admissão, laudos (laboratório de análise clínica, radiologia, eletrofisiologia, etc.), possíveis cirurgias e complicações, etc. Finalmente, os sumários incluem recomendações (medicação, etc.) e planos para o seguimento do caso. A concretização do sumário é, normalmente, realizada sob os auspícios do Sistema de Informação Hospitalar, quando do registro da alta médica.

3.2 Amostras do estudo

Por não se tratar de norma regida por lei, no Brasil nem todas as unidades de saúde sistematizam a construção do sumário de alta quando do atendimento clínico. Dentre as unidades de saúde que adotam a elaboração do sumário de alta na sua estratégia de comunicação com os pacientes e transparência clínica, o Hospital das Clínicas de Porto Alegre - HCPA é um dos precursores, tendo adotado o sumário de alta em 1997 (Klück e Guimarães, 1999).

O HCPA é uma empresa pública de direito privado criada pela Lei 5.604, de 02/09/1970, sendo parte da rede de hospitais universitários do Ministério da Educação e é vinculado à Universidade Federal do Rio Grande do Sul - UFRGS. O início das atividades ambulatoriais ocorreu em 1972, em 1973 as internações foram iniciadas. A excelência do HCPA é reconhecida, sendo considerado como uma das unidades de saúde de referencia pelo Ministério da Saúde, contando, atualmente, com 750 leitos (HCPA, 2009).

A solicitação de um conjunto de sumários de alta (descaracterizados) foi realizada junto ao HCPA através do projeto CONEP FR – 126681. O parecer positivo do comitê de ética do HCPA permitiu o acesso a um *corpus* composto pela totalidade dos sumários de alta de 5 anos da área de cardiologia e de 1 mês dos sumários de alta referente a todos os atendimentos hospitalares, sendo formado por um conjunto de 8.332 documentos, sendo esta a base que foi utilizada para os experimentos da presente tese.

Objetivando validar o conjunto de documentos obtidos junto ao HPCA, em especial na dimensão de vícios de amostra (*bias*), dois *corpora* de texto foram considerados:

1. Mac-Morpho *Corpus*, formado pela seleção randômica de textos do jornal Folha de São Paulo, do ano de 1994, sendo composto por 1.1 milhão de *tokens* (Aluisio, Pinheiro e Finger, 2003)*;* e
2. provenientes dos atendimentos ambulatoriais dos Hospitais Cajuru e Santa Casa de Misericórdia, entre os anos de 2007 e 2009, 38.421 anamneses que compõem o *corpus* Aliança.

Na Figura 14 é apresentada uma representação gráfica dos *corpora* utilizados nos experimentos da tese.

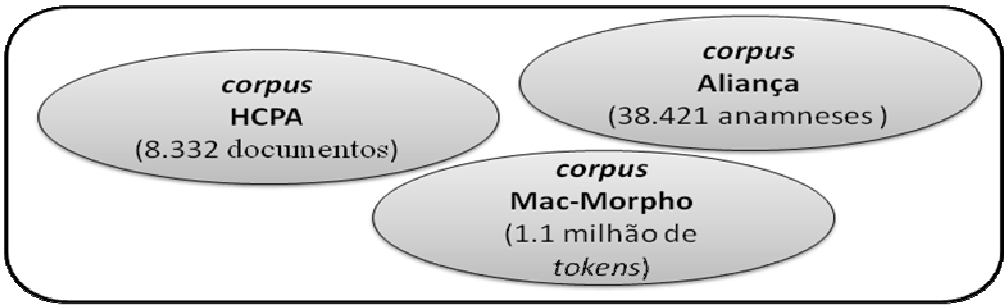


Figura 14 — Representação gráfica detalhada dos *corpus* que foram utilizados no desenvolvimento da tese

A Aliança Saúde (Aliança Saúde, 2009) está diretamente relacionada à Pontifícia Universidade Católica do Paraná e realiza, aproximadamente, 720 internações e 32.000 consultas ambulatoriais por mês, sendo que 75% dos atendimentos são realizados através do SUS.

Em processo de informatização recente, a Aliança Saúde não inclui na sua rotina de atendimento a produção do sumário de alta, sendo que os ambulatórios dos Hospitais Cajuru e Santa Casa de Misericórdia foram informatizados em meados de 2008. Neste sentido, optou-se pelo uso de anamneses, dado sua semelhança estrutural com os sumários de alta. O acesso aos textos ocorreu por meio do projeto CONEP FR – 274810. A base das 38421 anamneses ambulatoriais contempla o conjunto domínio disponível até 5 de janeiro de 2009. Cabe

ressaltar, no entanto, que apesar das semelhanças em estrutura, o texto contido nas anamneses são, geralmente, pouco gerais, normalmente focados no atendimento do paciente e, desta forma, foram utilizados apenas para a etapa de avaliação e validação do *corpus* HCPA.

Com base no *corpus* HCPA foi possível identificar 25.027 *tokens* únicos (palavras, acrônimos, numerais, entre outros), de um total de 861.596 *tokens*, com média de 103,40 *tokens* por prontuário e desvio padrão de 25,3. Na Tabela 3, apresenta-se uma amostra de *tokens* extraídos, cabe destacar que, a exemplo da palavra 'medocações', existem ocorrências de *tokens* incorretamente grafados.

Tabela 3 — Amostra de *tokens* extraído do conjunto de sumários de alta disponibilizados pelo HCPA (5 anos da área de cardiologia e de 1 mês dos sumários de alta referente a todos os atendimentos hospitalares)

posição relativa	***token***	**número de ocorrências no *corpus***
1ª	De	195904
10ª	Paciente	6372
18ª	miligrama	2928
22ª	Artéria	2597
24ª	Coronária	2467
102ª	Arteria	872
375ª	250mg	261
453ª	Ic	203
454ª	Ig	203
1821ª	Cisto	34
2120ª	Herpes	27
12949ª	fe25	10
25024ª	medocações	1

O Mac-Morpho *Corpus* é parte do projeto LacioWeb e está disponível para a comunidade científica mediante solicitação de autorização de uso (LácioWeb, 2009). O *corpus* em questão foi utilizado para a avaliação dos *tokens* presentes nos sumários de alta avaliados.

Com base no *corpus* da Aliança Saúde, foi possível identificar 12.027 *tokens* únicos (palavras, acrônimos, numerais, entre outros), de um total de 6.140.188 *tokens*, com média de 159,81 *tokens* por anamnese e desvio padrão de 60,6.

A relação cruzada dos *tokens* únicos, tendo como base o *corpus* HCPA, foi determinada através da análise comparativa entre os textos, sendo o conjunto interseção entre o *corpus* HCPA e Mac-Morpho (Figura 15 - ∩) é formado por 54,3% dos *tokens* únicos do *corpus* HCPA.

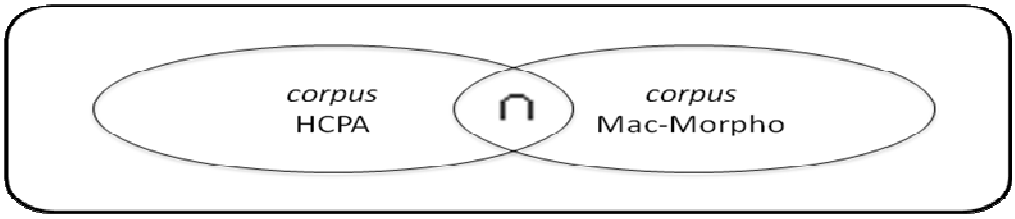


Figura 15 — Untersecção formada pelo conjunto de *tokens* únicos entre o corpus HCPA e Mac-Morpho

Na mesma direção, o conjunto de *tokens* únicos comuns entre o *corpus* do HCPA e Aliança Saúde (Figura 16 - ∩) é formado por 91,7% dos *tokens* únicos do *corpus* do HCPA.

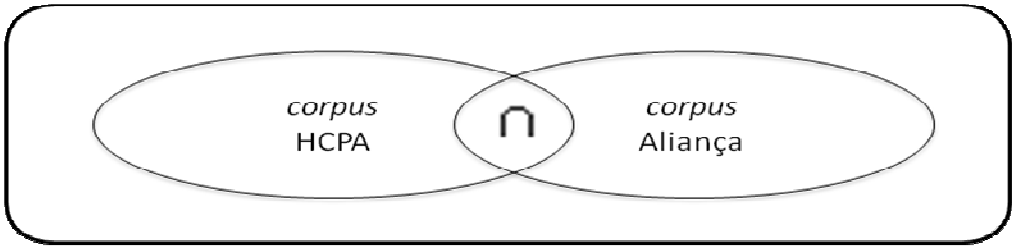


Figura 16 — Intersecção formada pelo conjunto de *tokens* únicos entre o corpus HCPA e Aliança

3.3 Etapas do estudo e bases para o experimento

A etapa inicial do trabalho consistiu do processamento de linguagem natural aplicado ao conjunto de narrativas clínicas do *corpus* do HCPA. Para a consecução desta atividade, detalhada no Capítulo 4, as seguintes ações foram realizadas:

a) levantamento de fenômenos lingüísticos: que objetivou a determinação das características lingüísticas do *corpus*. A avaliação aconteceu com base no conjunto C2, conforme representado na Figura 17;
b) correção ortográfica: que diminui o impacto de erros no PLN. A preparação da base de termos foi realizada com base em C3 (Figura 17);
c) normalização ortográfica: que contempla a expansão de acrônimos e resolução de abreviações. A base de dados para a concretização da atividade foi realizada com base em C2 (Figura 17);

d) etiquetagem morfológica: para a etiquetagem morfológica foi constituída uma base de treinamento, nominada de TREINAMENTO, e validação, nominada de VALIDAÇÃO, conforme apresentado na Figura 17, sendo adotada a metodologia de *Active Learning*[14] para a consecução da mesma; e

e) identificação de candidatos a frases nominais: etapa que consiste na identificação de frases nominais em uma sentença, processo validado com base no conjunto VALIDAÇÃO.

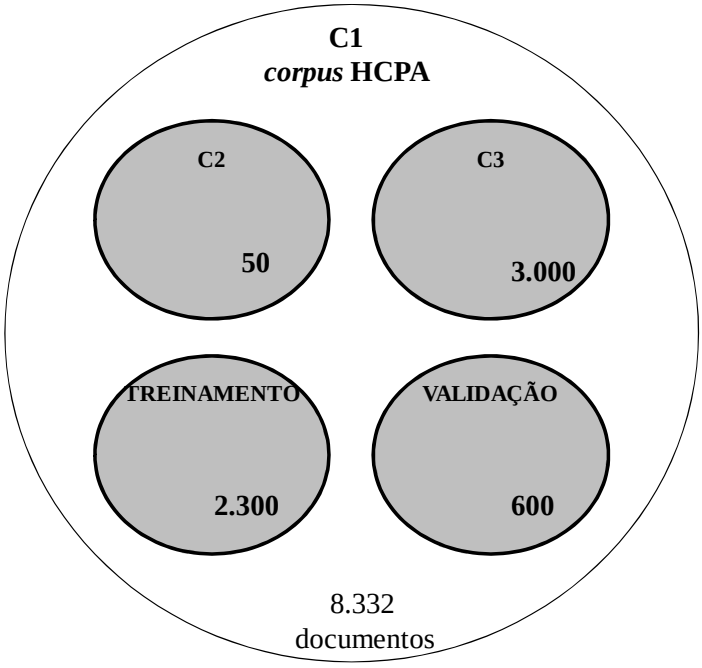


Figura 17 — Representação dos conjuntos de documentos, com suas respectivas quantidades, utilizadas na etapa de processamento de linguagem natural

Face às inúmeras variações lingüísticas, característica da linguagem natural, é necessária, quando da aplicação de ferramentas de PLN, a adoção de estratégias para o controle de vocabulário (Jacquemin e Klavans, 1997). Para o alcance deste fim, adotou-se a normalização lingüística.

As técnicas tradicionais para a normalização lingüística não permitem um eficiente controle de vocabulário para o domínio médico, conforme detalhado no Capítulo V, neste cenário, utilizou-se o sistema MorphoSaurus, por contrapor técnicas de normalização e um tesauro de domínio.

Para a adequação das ferramentas e do tesauro do sistema MorphoSaurus ao propósito da tese, as seguintes ações, detalhadas no Capítulo V, foram realizadas:

[14] Metodologia utilizada para o treinamento de ferramentas de etiquetagem morfológica com base em um comitê de indicadores, que guiam o processo de correção e avaliação de sentenças.

a) resolução de ramificações e ciclos: em função das características construtivas do tesauros do MorphoSaurus, muitos problemas nas relações levavam a ciclos, impactando diretamente no desempenho e na qualidade dos resultados;
b) detecção de idiossincrasias baseada na comparação de corpora similares: o tesauro do MorphoSaurus é multilíngüe, neste sentido, para sua constituição, equipes geograficamente separadas desenvolveram o tesauro ao longo dos anos. Como resultado deste fluxo de trabalho, problemas de delimitação de termos entre idiomas, tornou-se comum na estrutura do tesauro, demandando ação para a aplicação da tecnologia no escopo da presente tese; e
c) validação de alterações baseado em *logs*: complementar a detecção de idiossincrasias baseada na comparação de corpora similares, esta atividade consiste na identificação de inconsistências durante a etapa de desenvolvimento do tesauro.

Para o mapeamento das narrativas para uma ontologia médica foi selecionada a SNOMED CT, outras terminologias e ontologias foram consideradas, como a CID10, GALEN, MedO (OPENCLINICAL, 2009), no entanto, apenas a SNOMED CT apresenta estrutura suficientemente abrangente e quantidade de conceitos necessária para a concretização do presente trabalho.

O padrão ouro para o treinamento e validação do mapeamento de texto livre para a SNOMED CT foi construída com a inter-relação de quatro profissionais de domínio, sendo dois brasileiros e dois alemães. Como prerrogativa inicial, dada a complexidade do conhecimento codificado, como ilustrado na Figura 18, optou-se pela codificação com a construção de conceitos pós-coordenados. A estratégia de construção de conceitos pós-coordenados permite aumentar a cobertura do mapeamento, mas conduz a uma estratégia de maior complexidade e os resultados gerados são mais sensíveis à interpretação de cada pesquisador.

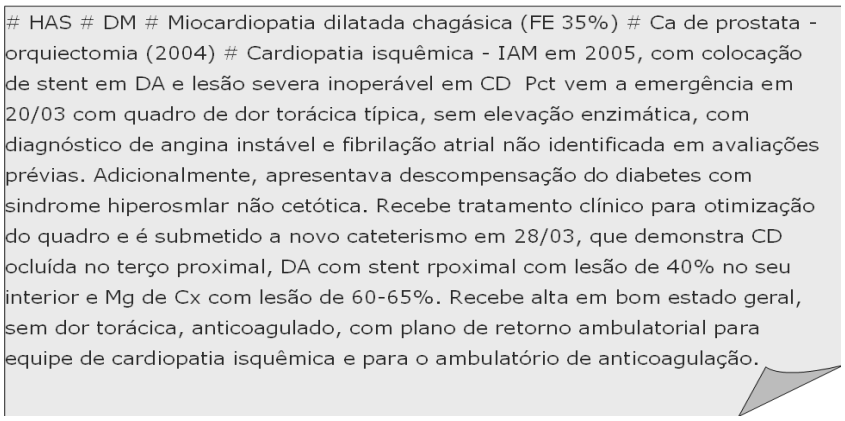

HAS # DM # Miocardiopatia dilatada chagásica (FE 35%) # Ca de prostata - orquiectomia (2004) # Cardiopatia isquêmica - IAM em 2005, com colocação de stent em DA e lesão severa inoperável em CD Pct vem a emergência em 20/03 com quadro de dor torácica típica, sem elevação enzimática, com diagnóstico de angina instável e fibrilação atrial não identificada em avaliações prévias. Adicionalmente, apresentava descompensação do diabetes com sindrome hiperosmlar não cetótica. Recebe tratamento clínico para otimização do quadro e é submetido a novo cateterismo em 28/03, que demonstra CD ocluída no terço proximal, DA com stent rpoximal com lesão de 40% no seu interior e Mg de Cx com lesão de 60-65%. Recebe alta em bom estado geral, sem dor torácica, anticoagulado, com plano de retorno ambulatorial para equipe de cardiopatia isquêmica e para o ambulatório de anticoagulação.

Figura 18 — Exemplo de sumário de alta utilizado

Após a análise de dez documentos, verificou-se que a estratégia inicialmente proposta não era adequada ao propósito da pesquisa. Neste sentido, adotou-se a codificação através de conceitos pré-coordenados (definidos *a priori* na base da SNOMED CT). Para os casos complexos, em que um único conceito não é suficiente para representar os termos a serem mapeados, definiu-se pela utilização de mais de um conceito para o mapeamento dos termos.

Para o estabelecimento do padrão ouro para o mapeamento, foi constituído um g*uideline* de trabalho, baseado nos trabalho de (Andrews, Richesson e Krischer, 2007; Lussier, Shagina e Friedman, 2001; Patrick, 2008; Andrews, 2008; Long, 2007), com as diretrizes de ação, objetivando guiar a anotação manual de sumários de alta para conceitos da SNOMED CT.

A anotação semântica consiste da análise estrutural de uma sentença, a identificação de coordenações (ou termos), a atribuição de marcadores de conceito e, finalmente, a atribuição de 0..n conceitos SNOMED CT por termo.

Para cada sumário de alta analisado, construiu-se um documento estruturado com as seguintes seções:

- *Chunk*: equivalente a uma sentença nominal ou, ainda, termo;
- *Polarity:* indica se o termo indica negação (NEG) ou não;
- *Code:* conjunto de códigos ativos na SNOMED CT, e
- *Viewpoint*: a interpretação de um texto está diretamente relacionada a elementos que remetem à categoria gramatical, como o contexto, aspectos, dentre outros. Como simplificação deste cenário, para a análise posta, foi introduzida a notação de *viewpoint.* Os mais importantes *viewpoints* identificados nos sumários de alta, são apresentados na Tabela 4.

As seguintes regras foram estabelecidas no transcurso do trabalho, objetivando o estabelecimento de um padrão único de codificação entre os especialistas envolvidos:

- optar por conceitos similares ao invés de conceitos pós-coordenados;
- a codificação deve considerar apenas fatos relevantes (para o contexto da codificação e Recuperação de Informação aplicado a documentos clínicos);
- a codificação de verbos deve ocorrer somente nos casos dos mesmos representarem conceitos clínicos relevantes;
- não utilizar conceitos que façam parte da hierarquia "*context dependent categories*" da SNOMED CT, por não representarem informação relevante para o domínio médico;

- não utilizar conceitos que expressem negação, pois os mesmos correspondem a um erro ontológico e devem ser revistos nas próximas versões da SNOMED CT; ao invés utilizar *Polarity=NEG*;

Tabela 4 — Conjunto domínio utilizado para o *ViewPoint*

Viewpoint	Abrev.	Descrição
stopped	STP	refere-se a períodos de tratamento, como em: medicação foi suspensa, ou no final da observância de sinais ou sintomas (remissão)
history	HIS	usado estritamente nas condições de existência ou ocorrência anterior ao período de tratamento ("histórico de sintomas"). Exemplo: "paciente apresentou acidente vascular há dois anos.."
family history	FAM	refere-se a fatos relacionados a família do paciente ("pai apresenta problemas cardíacos")
uncertainty	UNC	refere-se a uma expressão probabilística ("disfunção respiratória provavelmente causada por...") ou ainda negação ("desconhecido", "sem prévia história de...")
plan	PLA	refere-se ao plano de uma intervenção ("a implantação de um *stent* foi planejada")
imperative/optative	IMP	refere-se a ações incertas/recomendações ainda não validadas ("um *stent* deverá ser implantado")
side effects	SEF	refere-se ao efeito não desejado de uma droga ou terapia
necessity	NEC	um procedimento é necessário, independente se o mesmo foi ou será realizado
risk	RSK	refere-se ao risco de um sintoma ou doença

- utilizar o conceito da hierarquia '*substance*' para drogas (ex. "*digoxin (substance)*" ao invés de "*digoxin (product)*"), para o caso de marcas, o mapeamento deve ser realizado para o nome do sal (genérico);
- utilizar sempre conceitos que estejam completamente definidos (*fully defined*);
- utilizar sempre o conceito mais específico possível, e
- utilizar códigos da hierarquia *morphology* somente quando não há código apropriado na hierarquia *finding code*.

Na Tabela 5 ilustra-se o mapeamento manual de um sumário de alta.

O mapeamento manual foi realizado sobre um conjunto de 200 prontuários, sendo 120 do conjunto TREINAMENTO e 80 do conjunto VALIDACAO (os prontuários selecionados foram submetidos à atividade de pré-processamento: correção ortográfica e expansão de acrônimos).

Tabela 5 – Sumário de alta manualmente anotado

Chunk	**Viewpoint**	**Polarity**	**Code**
Paciente			
com angina pectoris classe III			85284003
com lesão crítica em primeira marginal			22765000, 24484000, 233970002
interna			
para realização			
de angioplastia	PLA		41339005
e implante de stent	PLA		36969009
na referida lesão.	PLA		
Procedimento			
realizado			
em 10/06,			
resultou			
em resolução da lesão			
com fluxo angiográfico TIMI III.			371865008
Houve			
durante o procedimento			
pequena dissecção distal a lesão,			70390005, 255604002
sem comprometimento do fluxo coronariano.			301121007
Recebe			
alta em bom estado geral.			

3.4 Aspectos Tecnológicos

Para o presente trabalho interessam os processos que estão diretamente relacionados com transformações de dados de entrada em dados de domínio e não em processos relativos a aspectos operacionais, que realizam um conjunto de atividades em função da arquitetura operacional adotada (como comunicação, operações de persistência, validação de interface gráfica, entre outros). Esses processos, denotativos do conhecimento representado nos Sistemas de Informação, encontram-se distribuídos em diversos multimeios, demandando a adoção de estratégias de mapeamento e processamento de conhecimento, bem como a distribuição e desenvolvimento sob a forma de estruturas inter-dependentes, como os preconizados pela IA para o casos dos agentes de software.

De forma a considerar a percepção e transformação fenomenológicas, características da diversidade e variabilidade (técnica e metodológica) de um ambiente hospitalar, optou-se por entender o mapeamento como resultado da aplicação de processos sobre o modelo conceitual da aplicação.

A observação fenomenológica, característica das inter-relações no mundo sensível, é uma atividade essencialmente paralela, em função da interdependência entre os fenômenos. É factível descrever que os sentidos são “sensores” distribuídos enviando sinais para uma unidade central (cérebro).

Para a representação computacional de tal comportamento, optou-se por uma sociedade de agentes, baseado no *framework* UIMA, desenvolvido em Java, que permite um alto de grau de paralelismo, além da distribuição de tarefas entre diferentes unidades computacionais (quando em uma estrutura em rede). Outro fator preponderante na seleção da metodologia em questão reside na distribuição dos processos que referenciem situações de cognição entre diferentes entes (agentes).

Neste cenário, os agentes que compõem a sociedade são apresentados na Figura 19.

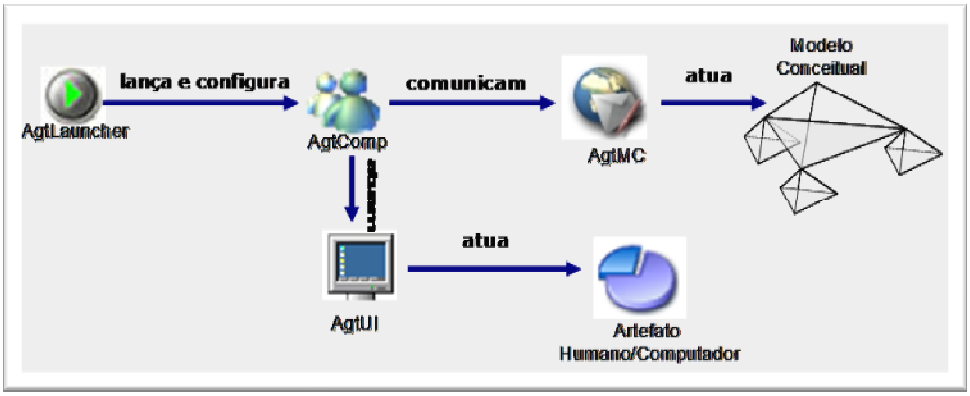


Figura 19 — Diagrama funcional da sociedade de agentes.

Sobre a Figura 19, é importante destacar:

a) o **AgtLauncher (*Ls*)** é um agente responsável pelo lançamento da sociedade e da delimitação interna dos agentes **AgtComp**;

b) os agentes **AgtComp** (autônomos e móveis), apresentam um conjunto de competências que os habilitam a, de forma individual ou coletiva, atuarem no Modelo Conceitual (no caso, o repositório de documentos e suas referências para a SNOMED CT);

c) o **AgtMC** recebe as mensagens de atualização do Modelo Conceitual, verifica a consistência e, então, efetiva a atualização do Modelo; e

d) o **AgtUI** recebe informações dos agentes **AgtComp** e as exibe em artefato próprio, permitindo o acompanhamento das ações dos respectivos agentes.

As competências dos agentes **AgtComp** podem abranger, de maneira combinada ou individualizada:

a) a **percepção fenomenológica,** que realiza a percepção de fenômenos do mundo sensível, com a aplicação de uma tecnologia específica (como o acesso a bases de dados ou *WebServices*);
b) a **atuação epistêmica e/ou crível**, que possibilita ao agente atuar nas heurísticas de mapeamento; e
c) a **transformação fenomenológica**, que permite a transformação processual de um fenômeno em outro fenômeno relevante para o mundo, através da aplicação de estruturas relativas à dedução ou indução.

A atuação no nível crível é baseada na metodologia de *Reforcement Learning* (Russel e Norvig, 1995) e, computacionalmente, contempla o reforço de uma associação quando da identificação que um fenômeno observado é classificado como positivo, em contrapartida, existe enfraquecimento de uma associação quando da classificação da observação como não positiva.

Buscando a flexibilização da ação dos agentes **AgtComp,** definiu-se um ambiente para a personalização da sociedade conforme apresentado na Figura 20, onde:

a) cada agente é identificado através de uma imagem (selecionada em função da respectiva competência) e de um texto delimitativo;
b) caso o agente atue na percepção ou transformação fenomenológica, torna-se necessário identificar os destinos do resultado da atuação do agente, realizado mediante a seta orientada “conectando” os agentes;
c) os fenômenos são caracterizados, computacionalmente, por um conjunto de dados. Nesse cenário, cada agente apresenta um conjunto, vazio ou não, de dados de entrada e de saída, que especificam os fenômenos “produzidos” ou “consumidos” pelo agente, e um conjunto de associações entre agentes de forma a caracterizar o destino dos fenômenos produzidos pela direta atuação das respectivas competências; e
d) são atualizados através da atuação do **AgtUI,** sendo que o item em questão exibe um conjunto de informações sobre o agente, seja de forma gráfica ou textual.

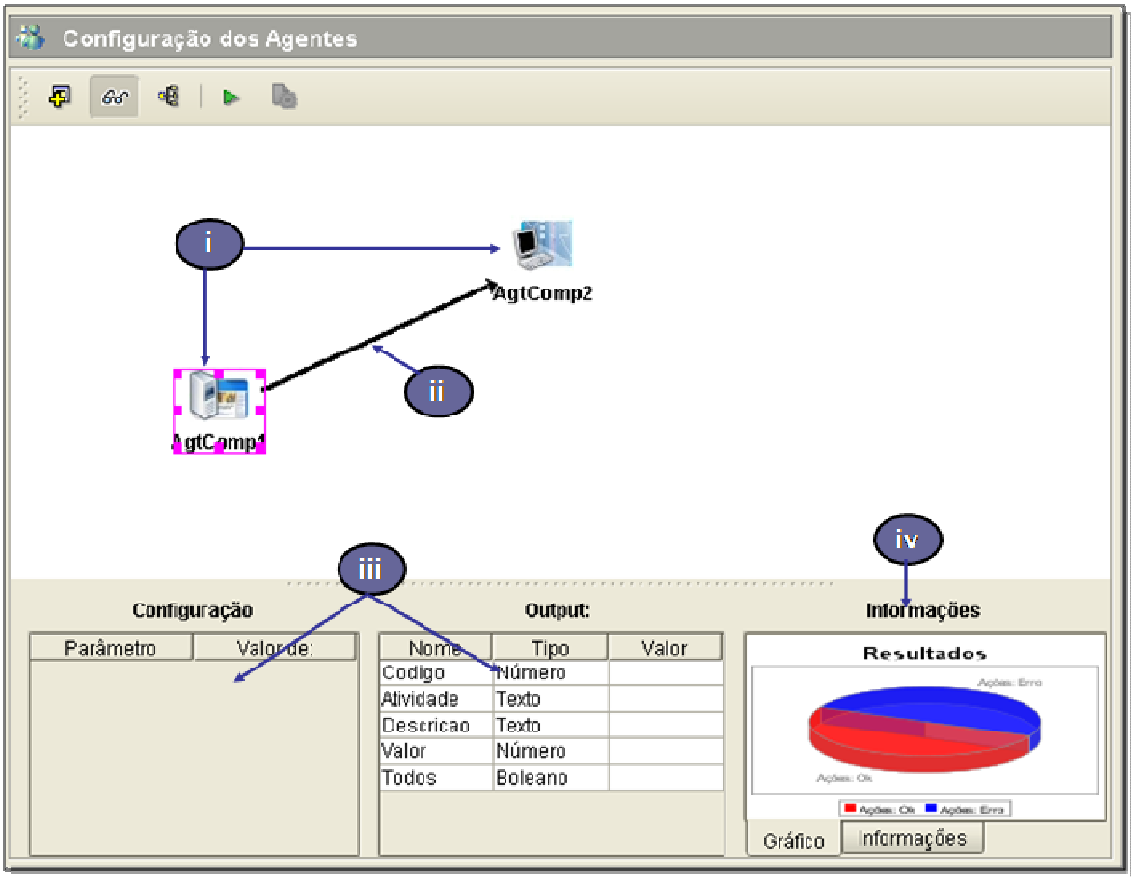


Figura 20 — Ambiente de Configuração dos Agentes

Sendo assim, um agente **AgtComp**, ao nível mais abstrato, é definido pela 4-upla ordenada **AgtComp**(**Comp, FenEnt, FenSai, Asso**) onde:

a) **Comp** é a competência do agente.

b) **FenEnt** é o conjunto (vazio ou não) de fenômenos observado pelo agente**.**

c) **FenSai** é o conjunto (vazio ou não) fenômenos produzido pela atuação da competência (Comp) do agente.

d) **Asso** é o conjunto de associações que indicam os destinos do fenômeno produzido.

Fazendo uso dos recursos de reflexão da linguagem Java, o ambiente permite a adição de novos agentes **AgtComp** e, por conseguinte, a adição de novas competências na sociedade.

De forma a propiciar infra-estrutura tecnológica para comunicação (em especial entre um **AgtComp** e os agentes **AgtMC** e **AgtUI)** e, também, para propiciar o formalismo contratual requisitado pelo ambiente de configuração, foi disponibilizada uma classe com o desenvolvimento básico de um **AgtComp**, servindo como um modelo para o desenvolvimento de novos agentes/competências, o diagrama UML respectivo é apresentado na Figura 21.

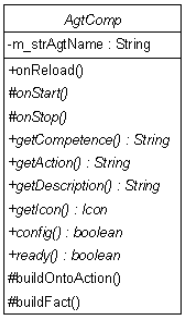


Figura 21 — Classe base para a especialização de novo agente **AgtComp**

A descrição dos métodos da classe AgtComp é destacada:

- **onReload** deve ser implementado na especialização dos agentes quando da necessidade da realização de algum procedimento especial na carga do agente, realizada pelo **AgtLauncher**;
- **onStart** e **onStop** são utilizados para o processo de sincronização com o mundo;
- **getCompetence, getAction, getDescription** e **getIcon** são responsáveis pela caracterização do agente no ambiente de configuração;
- **config** é chamado no momento de configuração do agente;
- **ready** é responsável pela determinação se, no instante considerado, o agente encontra-se pronto para ser lançado, e
- **buildOntoAction** e **buildFact** são utilizados no processo de depuração.

Finalizado o processo de implementação do novo agente, através da especialização dos métodos citados, a adição do agente no ambiente é realizada mediante a adição do arquivo *byte-code* (extensão .class).

Como resultado da arquitetura proposta e objetivando a realização de experimentos com o *framework* implementado, para a comprovação da funcionalidade do ambiente, um conjunto de agentes e competências foram implementados, em duas dimensões, quais sejam:

a) os necessários para o alcance do objetivo da presente tese; e
b) os que podem, oportunamente, serem utilizados, em conjunto ou isoladamente, na construção de novos experimentos (indicados como trabalhos futuros).

Na primeira dimensão os seguintes agentes foram desenvolvidos:

a) Percepção do *inner*.
b) PLN.
c) Seleção de NP.
d) MorphoMap.

A área de suporte e apoio à decisão poderá ser muito beneficiada com os resultados alcançados na presente tese. Por meio da utilização de estratégias conhecidas, é possível aliar maior qualidade e menores custos na assistência médica. Deste modo, na segunda dimensão, alguns agentes foram desenvolvidos, sem o objetivo de esgotar o tema, mas para permitir a construção de caminhos para a continuidade do projeto. Assim, os seguintes agentes foram desenvolvidos:

a) Busca de Definição;
b) Atuação crível através de Árvores de Decisão;
c) Transformação Fenomenológica baseada em Redes Baysianas;
d) Percepção Fenomenológica baseada em Banco de Dados;
e) Percepção Fenomenológica através de *WebServices*.

O agente "Percepção do *inner"*, que atua na dimensão de percepção fenomenológica, apresenta a capacidade de observação do mundo *inner* (Shmeil, 1999) (estrutura conceitual), detectando a presença de novos documentos a serem mapeados, sendo definido, ao nível mais abstrato, pela 4-upla ordenada AgtComp(Perceção do Inner, FenEnt, FenSai, Asso), onde:

a) Percepção do *Inner*: competência que viabiliza a observação do *inner* (estrutura conceitual), detectando a criação de novas instâncias ou conceitos, bem como a alteração dos mesmos;
b) FenEnt: não apresenta, de forma explícita, um fenômeno de entrada, já que o mesmo passa a ser observado pela atuação da competência do agente e não da comunicação entre agentes na sociedade;
c) FenSai: conceito ou instância observado no *inner*;
d) Asso: conjunto de associações que indicam os agentes, da sociedade configurada, que apresentam interesse na observação do agente especificado.

O agente "PLN" é responsável pelas atividades relacionadas ao processamento de linguagem natural, considerando as especificidades necessárias para a manipulação de narrativas clínicas (Capítulo 4) e engloba as etapas de:

a) Detecção de sentenças,
b) Verificação ortográfica,
c) Resolução de acrônimos,
d) NP *recognition*,
e) *POS Tagging*,
f) Extração de NPs.

Definido, ao nível mais abstrato, pela 3-upla ordenada AgtComp(FenEnt, FenSai, Asso), onde:

a) **FenEnt**: fenômeno de entrada, normalmente uma narrativa clínica;
b) **FenSai**: elementos identificados (sentenças identificadas com classes gramaticais, com meta-anotações delimitando as NPs.);
c) **Asso**: conjunto de associações que indicam os agentes, da sociedade configurada, que apresentam interesse na observação do agente especificado.

O agente de "Seleção de NP", com base nos critérios enunciados nos Capítulos 5 e 6, identifica os melhores candidatos a NPs para o contexto de aplicação, sendo definido, ao nível mais abstrato, pela 3-upla ordenada AgtComp(FenEnt, FenSai, Asso) onde:

a) **FenEnt**: exige a conexão com o agente NP ou outro que produza os mesmos objetivos serializados gerado pelo referido;
b) **FenSai**: NPs identificadas;
c) **Asso**: conjunto de associações que indicam os agentes, da sociedade configurada, que apresentam interesse na observação do agente especificado.

O agente "MorphoMap", com base em um conjunto de NPs, realiza o mapeamento das mesmas para a SNOMED CT, de acordo com as características enunciadas no Capítulo 6, sendo definido, ao nível mais abstrato, pela 2-upla ordenada AgtComp(FenEnt, Asso) onde:

a) **FenEnt**: conjunto de NPs que serão mapeadas para a SNOMED CT;
b) **Asso**: conjunto de associações que indicam os agentes, da sociedade configurada, que apresentam interesse na observação do agente especificado.

O agente de "Busca de Definição", por sua vez, realiza a busca do significado de um conceito em uma base estruturada, atuando na dimensão de percepção fenomenológica, sendo definido, ao nível mais abstrato, pela 4-upla ordenada AgtComp(Busca de Definição, FenEnt, FenSai, Asso) onde:

a) Busca de Definição: competência associada com a busca da definição de um conceito em multimeios, através de atuação a nível epistêmico;
b) FentEnt: conjunto composto do conceito que sofrerá a atuação do agente e do significado do identificador, obtido através da atuação de outro agente;
c) FenSai: identificador do conceito, objetivando a atuação de outro agente AgtComp que tenha competência associada a busca de definição de um termo;

d) Asso: conjunto de associações que indicam os agentes (na sociedade configurada) que apresentam interesse na observação do agente especificado.

A atuação padrão do **AgtComp** "Busca de Definição" é ilustrado na Figura 22:

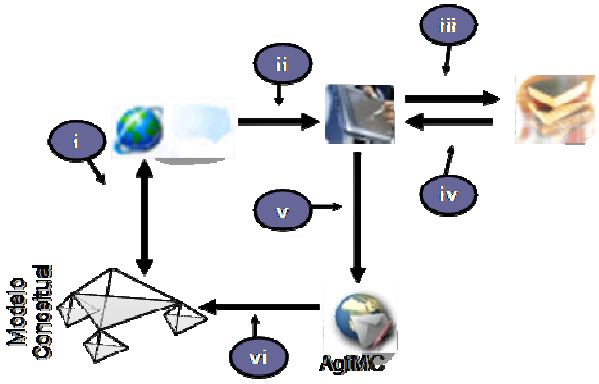


Figura 22 — Fluxo de atuação do agente "Busca de Definição"

i. o agente em questão recebe um conceito observado por outro agente (normalmente do agente de Percepção do Inner);
ii. seqüencialmente, realiza a extração do identificar de conceito;
iii. repassando-o para outro agente que tenha a competência associada à busca de definição em multimeios (ex.: busca de descrição na SNOMED CT);
iv. aguarda a mensagem contendo o significado do termo;
v. ao recebê-lo envia uma mensagem para o **AgtMC**; e
vi. o **AgtMC**, após validar o processo, atua no modelo conceitual.

O agente de "Atuação crível através de Árvores de Decisão" atua no fortalecimento ou enfraquecimento de associações baseado em fenômenos observados e é definido, ao nível mais abstrato, pela n-upla ordenada AgtComp (Atuação crível através de Árvores de Decisão, FenEnt, FenSai, Asso), onde:

a) Atuação crível através de Árvores de Decisão*:* realiza o fortalecimento ou enfraquecimento de associações, atuação baseada em árvores de decisão tendo como elementos de tomada de decisão os fenômenos recebidos;
b) FentEnt: fenômenos a serem considerados no processo;
c) FenSai: descritivo da ação realizada;

d) Asso*:* conjunto de associações que indicam os agentes (na sociedade configurada) que apresentam interesse na observação do agente especificado.

O processo de configuração da atuação do agente consiste na definição da árvore de que objetiva o fortalecimento ou enfraquecimento de associações, a decisão é configurável e baseada na percepção fenomenológica. Para tanto, a árvore de decisão é composta de nós de decisão (**Figura 23** item **i**) e de ação (**Figura 23** item **ii**).

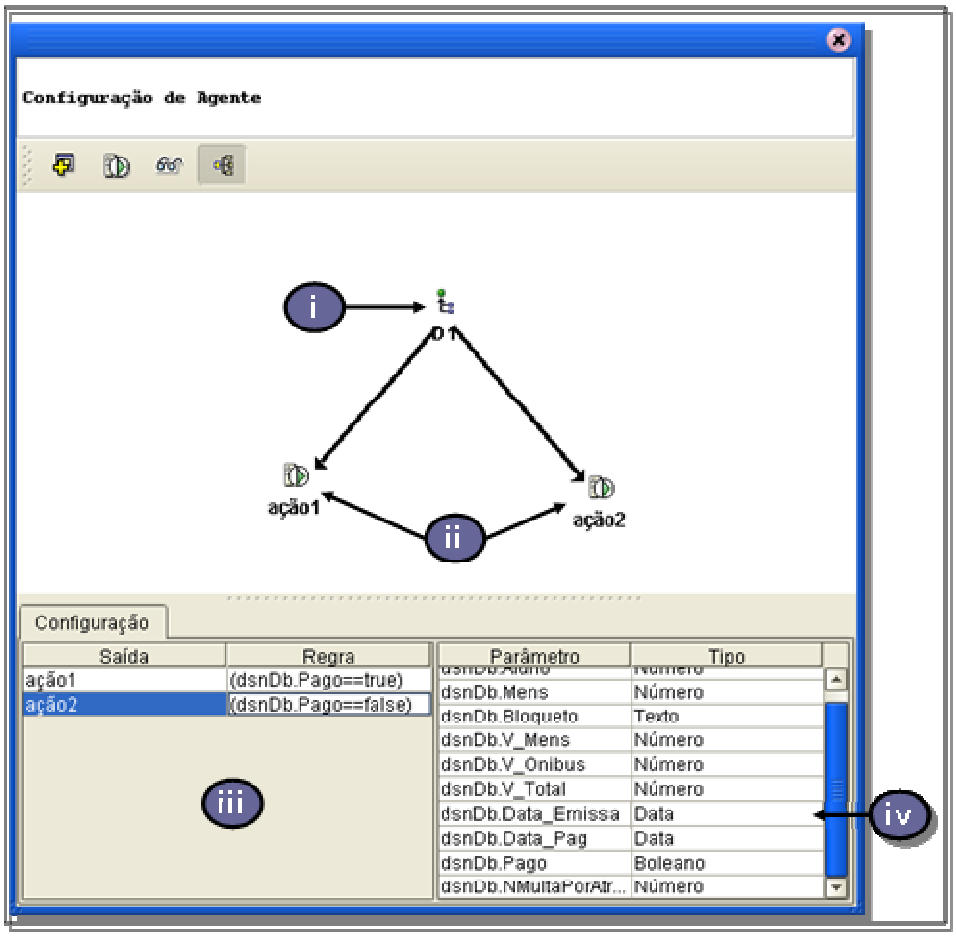


Figura 23 — Configuração de árvore de decisão - nó de decisão

Os nós de decisão realizam a seleção de qual dos elementos da árvore serão selecionados considerando o estado fenomenológico do momento da atuação (ação realizada através da avaliação de expressões booleanas - **Figura 23** item **iii**). De forma a auxiliar o processo de configuração são apresentados os componentes que compõem os fenômenos observados (**Figura 23** item **iv**).

O agente de “Transformação fenomenológica baseada em Redes Bayesianas”, atua em nível de transformação fenomenológica, calculando probabilidades considerando fenômenos de entrada e aplicando a teoria de Bayes, sendo definido, ao nível mais abstrato, pela 4-upla

ordenada AgtComp(Transformação Fenomenológica baseada em Redes Bayesianas, FenEnt, FenSai, Asso) onde:

a) **Transformação Fenomenológica baseada em Redes Bayesianas:** determina probabilidades, considerando a observação fenomenológica, através da aplicação da teoria de Bayes;
b) **FentEnt**: fenômenos a serem considerados no processo;
c) **FenSai**: resultado da atuação do agente;
d) **Asso:** conjunto de associações que indicam os agentes (na sociedade configurada) que apresentam interesse na observação do agente especificado.

O processo de configuração passa pela personalização dos agentes que fornecem os fenômenos para o processo considerado.

Durante o processo é possível configurar nodos que representam fenômenos observados por um agente do sistema, Figura 24, (sendo atualizados conforme a atuação do mesmo) e nodos com valores personalizados e fixos, Figura 25, utilizados para a especialização contextual.

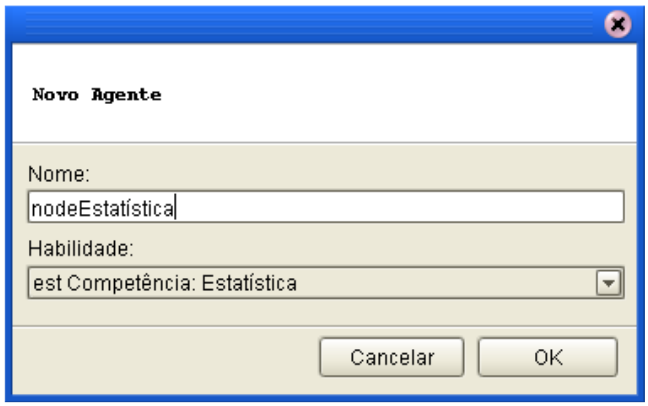


Figura 24 — Nodo associado com a observação fenomenológica

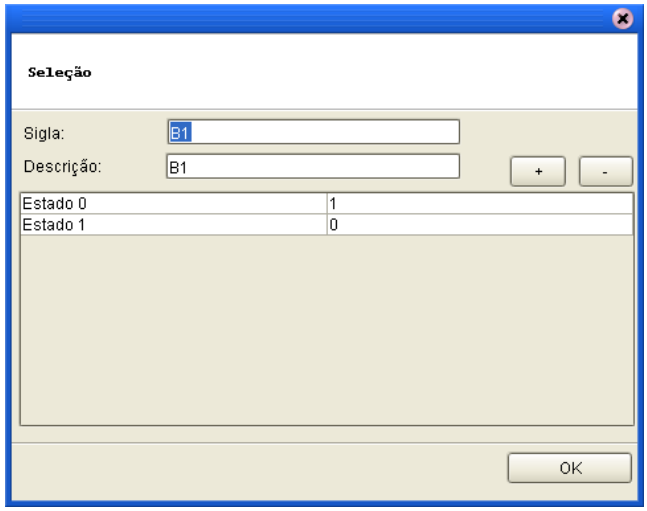


Figura 25 — Configuração de Nodo

Para a explicitação do nodo que representa o fenômeno resultante da atuação da competência do agente, utiliza-se um losango conectado ao nodo de probabilidade desejado, conforme apresentado na Figura 26.

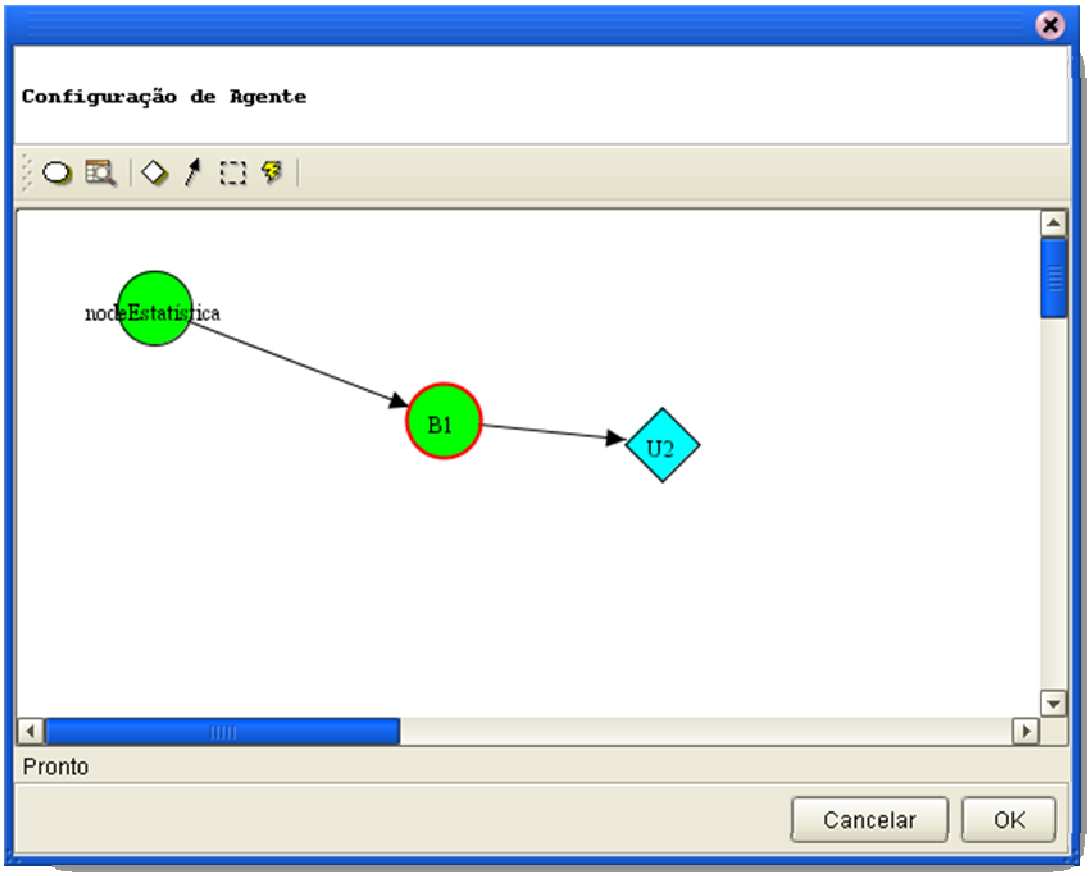


Figura 26 — Artefato de edição de rede bayesiana

O fenômeno resultante da atuação é composto dos elementos nomeados para o cálculo probabilísticos e os respectivos valores numéricos.

O agente “Percepção Fenomenológica baseada em Banco de Dados” permite o interfaceamento das demais competências com elementos recuperados de uma base de dados relacional, sendo definido, ao nível mais abstrato, pela 4-upla ordenada AgtComp(Percepção Fenomenológica baseada em Banco de Dados, FenEnt, FenSai, Asso) onde:

a) **Percepção Fenomenológica baseada em Banco de Dados:** atuação de percepção fenomenológica, objetivando a recuperação de fenômenos em bases relacionais,
b) **FentEnt**: não apresenta, explicitamente, nenhum fenômeno de entrada,
c) **FenSai**: fenômeno recuperado (caracterizado por seus elementos formadores), e
d) **Asso:** conjunto de associações que indicam os agentes (na sociedade configurada) que apresentam interesse na observação do agente especificado.

Por fim, o agente “Percepção Fenomenológica baseada em *WebServices*” permite a integração das demais competências com dados recuperados através de *WebServices*, sendo definido, ao nível mais abstrato, pela 4-upla ordenada AgtComp (Percepção Fenomenológica baseada em WebServices, FenEnt, FenSai, Asso) onde:

a) **Percepção Fenomenológica baseada em WebServices:** atuação de percepção fenomenológica, objetivando a recuperação de fenômenos em *WebSerices*;
b) **FentEnt**: parâmetros necessários para a recuperação de fenômenos;
c) **FenSai**: fenômeno recuperado (caracterizado por seus elementos formadores);
d) **Asso:** conjunto de associações que indicam os agentes (na sociedade configurada) que apresentam interesse na observação do agente especificado.

A atuação do agente em questão é baseada no padrão *Web Services Description Language* - WSDL, que contém informações sobre a estrutura de acesso de um WebService específico.

Capítulo 4

4 PROCESSAMENTO DE LINGUAGEM NATURAL APLICADO AO CONJUNTO DE NARRATIVAS CLÍNICAS

4.1 Introdução

O PLN, no domínio médico, também conhecido como Processamento da Linguagem Médica - PLM, tem sido uma área de intensa pesquisa nos últimos 30 anos, tento como projeto pioneiro o "*Linguistic String Project*" (Sager, 1967). Ao longo das últimas décadas o esforço principal das pesquisas esteve concentrado na análise de literatura nas áreas de biologia e genômica (Zweigenbaum, 2008).

Zweigenbaum (2008) destaca o crescimento do interesse na aplicação de técnicas de PLM, propondo e testando novos métodos, disponibilizando novas rotinas, como as análises automatizadas para relatórios clínicos (por exemplo, o MedLEE (Columbia University, 2009)).

O ponto focal relevante na maioria dos trabalhos na área de PLM é a particularidade estrutural e metodológica dos textos no domínio em questão. Firedman *et al.* (2002) destacam o contributo conceitual de Zellig Harris na proposição da teoria das sublinguagens, definindo as bases para o eficiente processamento de linguagens 'técnicas' (Harris, 1991), como a médica. Ao definir que tais linguagens apresentam estrutura e regularidade, Zellig propõe que esses elementos podem ser identificadas em *corpus* do domínio, no entanto, diferente da teoria gramatical padrão, que define estruturas sintáticas bem formatadas, a teoria de sublinguagens incorpora informações semânticas específicas do domínio de análise e relações probabilísticas entre os elementos sintáticos, estabelecendo as bases para aproximações baseadas em probabilidades de relações e não baseadas em regras fixas, bem definidas.

No domínio dos documentos clínicos, especificamente no caso das narrativas, a utilização da linguagem específica do profissional é requerida, dado o contexto de utilização e construção (Zweigenbaum, 2008). Adicionalmente, Hahn e Wermter (2004) destacam que o processamento de documentos clínicos oferece um novo conjunto de desafios para a área de PLN como, por exemplo, a variedade de estruturas não encontradas em documentos jornalísticos (normalmente o domínio de aplicação mais comum) e a necessidade de aprofundado conhecimento *a priori* sobre o domínio médico, por fim, Zweigenbaum (2008) observa que os documentos clínicos apresentam diferentes sublinguagens, em função da multidisciplinaridade característica.

O objetivo principal da pesquisa na área de PLN é a identificação morfológica e compreensão da linguagem. Para o alcance destes objetivos muitos pesquisadores na área tem focado esforços no alcance de uma atividade intermediária, objetivando a compreensão de algumas das estruturas relacionadas à linguagem, sem, no entanto, a premissa de compreensão completa das estruturas envolvidas: etiquetagem morfológica.

A etiquetagem morfológica atribui a cada palavra em uma sentença (contexto) uma etiqueta morfológica (substantivo, verbo, etc), representando sua classe gramatical; permitindo o desenvolvimento de outras estratégias, como a da identificação de frases nominais (Manning e Schuetze, 1999), que são unidades semânticas atômicas, passíveis de serem tratadas por algoritmos computacionais (Girju, 2006).

4.2 Levantamento de fenômenos lingüísticos

Com base em C2 (Figura 17) três especialistas de domínio analisaram manualmente os documentos, objetivando identificar e classificar os documentos quanto à qualidade dos textos (identificando casos de erro, como em: “**Pcte** apresentou hematúria no final da cirurgia , **comdição** que se agravou no centro **d** terapia intensiva...”) e a presença de acrônimos e abreviações (como em: “Paciente com estenose mitral reumática severa, sintomas de ICC C II-III”).

A classificação adotada para os tipos de erros identificados é a apresentada por (Bustamante e Díaz, 2006) e abrange erros:

a) ortográficos;
b) gramaticais;
c) de pontuação, e
d) de digitação.

Formalmente, distinguem-se:

a) omissão: deleção ou remoção, no qual letra, palavra, espaço ou pontuação é excluído;
b) substituição ou troca: no qual uma letra ou palavra correta é substituída por uma incorreta; substitui-se caixa alta por caixa baixa ou o contrário;
c) inserção: com adição de símbolos, letras que não pertencem à palavra, adição de palavras em frases, e
d) inversão: com troca de letras adjacentes ou não adjacentes em uma mesma palavra, ou ainda mais de um fenômeno em apenas uma palavra (Ren e Perrault, 1992), e capitalização: escrita de todas as palavras em letras maiúsculas.

Os resultados obtidos são agrupados nas doze dimensões:

a) Alguns **erros ortográficos** ocorrem devido às palavras homófonas (mesma pronúncia, diferente escrita, como em: mau e mal, conserto e concerto, sela e cela (Vosse, 1992)). Comumente, trocam-se fonemas com pronúncia semelhantes, como S, Z, Ç, S, G, J, U, L. Como em “fraquesa”, (o correto é fraqueza). Em 92% dos textos analisados, foram identificados casos de erros ortográficos;
b) Os **erros gramaticais** foram observados em 88% dos documentos analisados. Exemplos: **omissão de palavras**: “Após procedimento déficit neurológico - diplopia + Ataxia. RNM de crânio. Exame realizado em...” (ausência de verbos, o que caracterizaria um período); e **substituição de palavras**: “encaminhada de Erechim para cateterismo cardíaco ***para*** valvulopatia...” (substituição de **e** por **para**);
c) Os **erros de pontuação** estão presente em 96% dos textos de C2, como no exemplo “...MELHORA DO LEUCOGRAMA RECEBE ALTA EM BOM ESTADO...” (omissão de ponto), e
d) Os **erros de digitação** estão presente em 92% dos textos de C2 e, normalmente, resultam em omissão, substituição, capitalização, inserção, duplicação de espaços. Ao invés de erros gramaticais (onde o mesmo autor tende a repetir os mesmos erros) esses lapsos produzem resultados aleatórios. Exemplos:
e) Omissão de letras: “abdomina”, “tard”; “propanolol”;
f) Omissão de espaços. “...coronária direita.Procedimento...”,Paciente diabético,hipertenso...”; “ticlopidina?contraste?infecção?“;

g) Capitalização.“...PERDA TRANSITÓRIA DA CONSISTÊNCIA...”; “...TERÇO MÉDIO DA...”;

h) Trocas: “hipertemsão”, “leucograna”, geralmente por letras que correspondem a teclas adjacentes;

i) Inserções: “eviolução”, revelando o mesmo fenômeno;

j) Acentuação: “elevacao”;

k) Excesso de espaços. “Doença primariamente pulmonar...”, e

l) Inserção de símbolos e omissão de espaço. “...#ACTP COM IMPLANTE...”.

Os especialistas de domínio, com base em C2, identificaram que todos (100%) os documentos analisados continham abreviações e acrônimos, com média de 3,1 (acrônimos ou abreviações) por documento. Em nenhum dos documentos encontrou-se o par acrônimo e expansão, invalidando a aplicação de estratégias tradicionais de normalização.

4.2.1. Correção Ortográfica

Considerando a elevada prevalência de erros nos documentos analisados, foi definida uma estratégia de quantificação de similaridade entre *tokens*, possibilitando a correção ortográfica, com base na similaridade normalizada de Levenshtein (Yujian e Bo, 2007).

A estratégia de correção ortográfica é embasada nas evidências apontadas por Pollock e Zamora (1984), na qual erros, de quaisquer naturezas, ocorrem geralmente sem um padrão comum em grandes amostras; neste sentido, *tokens* com baixa freqüência são fortes candidatos a representarem um caso de erro (independente da categoria considerada).

Para a aplicação da similaridade normalizada de Levenshtein efetuou-se o levantamento do dicionário de *tokens* corretamente gravados no *corpus*. Para tanto, foi utilizado o conjunto C3 (Figura 17). Com base em C3 foi identificada a lista de *tokens* com mais de 4 caracteres (objetivando a eliminação de acrônimos e similares), compostos apenas por caracteres alfanuméricos, bem como sua freqüência no *corpus* e que ocorram ao menos em 10 documentos distintos. O dicionário levantado é composto de 8.921 *tokens*.

O algoritmo para correção ortográfica considerou grau de similaridade, com base d_{N-GLD}, igual ou superior a 80%.

A validação da estratégia foi realizada contra C2. 86% dos erros identificados foram corretamente corrigidos. 0,06% dos *tokens* foram indevidamente corrigidos.

4.2.2. Normalização

Segundo Pakhomov (2001), a “normalização de textos é um aspecto importante na aplicação de técnicas de PLN em documentos médicos”. Por normalização, entende-se a expansão de acrônimos e resolução de abreviações.

Nos documentos médicos, é comum a utilização de acrônimos e abreviações (Park, 2001). Neste domínio, a literatura apresenta estratégias para a detecção e expansão de acrônimos e abreviações, embasadas na evidência que, normalmente, nos documentos técnico-científicos, diferentemente das narrativas clínicas como identificado quando do “levantamento de fenômenos lingüísticos” na primeira ocorrência de um acrônimo a expansão co-existe na sentença (Torii, Liu e Hu, 2006; Ao e Takagi, 2005), como em “A unidade de terapia intensiva (UTI)...”.

Com base nos resultados auferidos na análise dos documentos que compõem C2, e, parcialmente, nas expressões regulares propostas no algoritmo ALICE para a extração de abreviações do MEDLINE (Ao e Takagi, 2005), um conjunto de regras foi desenvolvido e validado contra a análise manual realizada, e as premissas, codificadas sobre a forma de expressões regulares, para identificação de candidatos são apresentadas:

a) não contém espaço (objetivando descartar casos como “A B”);
b) comprimento do *token* analisado deve ser maior do que 1 (objetivando descartar casos como o de “A”);
c) no caso do primeiro caractere ser alfabético, deve existir pelo menos um caractere no *token* que seja capitalizado (objetivando a identificação dos casos como “mRNA” e “GmbH”); e
d) no caso do primeiro caractere ser numérico, deve existir pelo menos um outro caractere alfabético (excluindo os casos de “33” , “3,2” e “1/2/1”; mas identifica “5-fu”).

A aplicação das premissas citadas, em C2, possibilitou a identificação de 153 dos 155 acrônimos manualmente apontados, bem como outros 34 falsos positivos.

Objetivando a normalização dos documentos, realizou-se o levantamento do conjunto de acrônimos candidatos em C1, assim como a quantidade de ocorrências no conjunto referenciado, conforme ilustrado na Tabela 6.

Com base no conjunto de candidatos identificados os especialistas de domínio expandiram manualmente a lista gerada, para candidatos com ocorrência superior a 10. Desta análise, foi construído um dicionário de acrônimos (aqui nominado de **D**) e sua expansão. Os casos de falso positivo também foram anotados, conforme ilustrado na Tabela 6.

Tabela 6 — Amostra de acrônimos candidatos com seu respectivo significado no domínio analisado, extraídos de C1

posição relativa	candidato	número de ocorrências no *corpus*	significado principal no domínio
1ª	DE	1320	**FALSO POSITIVO**
2ª	AAS	966	ácido acetil salicílico
3ª	ACTP	810	angioplastia coronária transluminal percutânea
4ª	HCPA	772	Hospital de Clínicas de Porto Alegre
12ª	IAM	470	infarto agudo do miocárdio
16ª	TC	412	tomografia computadorizada
35ª	ACD	161	artéria coronária direita
63ª	MP	95	Marcapasso
117ª	HPS	44	Hospital de Pronto Socorro
148ª	e/ou	34	**FALSO POSITIVO**
167ª	PLANO	28	**FALSO POSITIVO**
171ª	MSE	27	membro superior esquerdo
222ª	QRS	18	repetição de complexos do eletrocardiograma

No total, foram identificados 314 acrônimos candidatos e, destes, 11,78% (37 candidatos) foram classificados como falsos positivos.

O dicionário obtido, conforme amostra da Tabela 6, foi validado com base nas 38.421 anamneses dos Hospitais da Aliança Saúde. A extração dos acrônimos candidatos nesta base permitiu identificar que 92,85% dos candidatos apontados (positivos ou não) fazem parte de D.

4.2.3. Etiquetagem morfológica: constituição da base de treinamento

Para o delineamento do domínio de base de treinamento para a etiquetagem morfológica, foi inicialmente estabelecido o conjunto mínimo de *tokens* necessários para o treinamento das ferramentas de PLN. Segundo Sardinha (2004), a literatura apresenta vários casos de etiquetadores que foram treinados em *corpora* que variavam de 20 mil (o mais comum) a 200 mil *tokens* (mais raro), sendo que há evidências de que o aumento na quantidade de *tokens* na etapa de treinamento auxilia no desempenho do etiquetador. Villavicencio e Viccari (1996) destacam a melhora no desempenho de um etiquetador morfológico, de 70 para 84,5% quando passaram de um *corpus* de treinamento de 700 para 13 mil *tokens*.

A etapa de treinamento de ferramentas de etiquetagem morfológica é uma atividade que demanda uma grande quantidade de recursos humanos, em face à necessidade da criação

de uma base manualmente etiquetada. Basicamente, a etiquetagem manual consiste no apontamento, para cada *token*, da sua etiqueta semântica. Nos experimentos iniciais realizados, um documento com 150 *tokens* demanda, em média, 15 min (trabalho em par) para a etiquetagem.

Para o processo em questão foi utilizado o conjunto TREINAMENTO, com 2300 documentos e 253825 *tokens*, e VALIDAÇÃO, com 600 documentos e 79587 *tokens*.

Os conjuntos TREINAMENTO e VALIDACAO passaram por processos automatizados de normalização e correção ortográfica, conforme apresentado em 4.2.1 (Correção Ortográfica) e 4.2.2 (Normalização).

O conjunto VALIDACAO foi manualmente[15] etiquetado com base nas etiquetas apresentadas na Tabela 7.

Tabela 7 — Etiquetas utilizadas

Classe Gramatical	**Etiqueta**
Pronome	PRN
Nome próprio	NPROP
Substantivo ou Adjetivo	NADJ
Numeral	NUM
Advérbio	ADV
Artigo	ART
Conjunção	CJ
Preposição	PREP
Palavra denotativa	PDEN
Particípio	PCP
Interjeição	IN
Verbo	V
Símbolo de moeda corrente	CUR

[15] O processo de etiquetagem manual foi realizado por 3 profissionais com conhecimento da área médica, 60% dos documentos foram co-validados.

4.2.4. Metodologia de treinamento

A metodologia de '*Active Learning*', comumente utilizada para o treinamento de classificadores, já para o treinamento de ferramentas de etiquetagem morfológica é um paradigma recente (Tomanek, Wermter e Hahn, 2007). Baseia-se no uso de um comitê de etiquetadores — cada qual utilizando métodos de análise específicos ou comuns — para avaliar as sentenças com maior grau de discordância e que, portanto, necessitam de avaliação manual.

Segundo Tomanek, Wermter e Hahn (2007), a estratégia é capaz de reduzir em até 50% o número de palavras a serem etiquetadas por humanos se comparada a uma técnica de seleção aleatória, alcançando os mesmos índices de exatidão.

Essa metodologia implica em duas decisões centrais, anteriores ao processo de desenvolvimento. A primeira consiste na escolha do comitê de etiquetadores. Nesse sentido, optou-se pelas ferramentas *Lácio-Web — MXPOST*[16], *TreeTagger*[17] e *Brill Tagger*[18] —, pelo etiquetador *QTag*[19] e pela *OpenNLP*. O comitê de etiquetadores foi selecionado considerando a exatidão mínima superior a 90%, quando aplicados em textos jornalísticos.

A segunda decisão importante refere-se ao conjunto de etiquetas que terá prioridade sobre os demais, para o qual todas as saídas devem ser mapeadas. O conjunto selecionado para o domínio de aplicação é apresentado na Tabela 7, com prioridade decrescente.

O estado inicial foi construído com base no *corpus* de TREINAMENTO e treinado com as ferramentas disponíveis na OpenNLP (OPENNLP, 2009), com base no modelo produzido a partir do Mac-Morpho *Corpus*[20]. O etiquetador morfológico da OpenNLP é baseado no algoritmo de Viterbi para modelos de Markov de segunda ordem. O *corpus* de TREINAMENTO devidamente etiquetado é nominado de T2.

A partir da definição do estado inicial, seguem-se iterações de treinamento com cada uma das ferramentas que compõem o comitê de etiquetadores, com base em T2. Para cada

---

[16] Baseada em um modelo de máxima entropia, *MXPOST* é um etiquetador desenvolvido por Adwait Ratnaparkhi do Departamento de Tecnologias em Lingüística Humana da IBM. Seu algoritmo de processamento baseia-se no uso de árvores de decisão, que implicam em estruturas condicionais.

[17] Desenvolvida pelo Instituto de Lingüística Computacional da Universidade de Stuttgart, *TreeTagger* é uma ferramenta probabilística de anotação de textos, que toma decisões através de uma árvore binária. Há registros de sucesso nos idiomas alemão, inglês, francês, italiano, espanhol e búlgaro (Ratnaparkhi, 1998).

[18] *Brill Tagger* é um etiquetador morfossintático baseado em regras elaborado por Eric Brill, na época professor da Universidade Johns Hopkins. Desenvolvido em *C*, seu algoritmo baseia-se na correção de etiquetas atribuídas anteriormente de acordo com a palavra atual (Ratnaparkhi, 1998).

[19] *QTag* é uma ferramenta de etiquetagem de textos desenvolvida pela Universidade de Birmingham no Reino Unido. Implementada em *Java*, consiste em um etiquetador de análise probabilística, que resolve ambigüidades por meio de ferramentas estatísticas (Ratnaparkhi, 1998). Para o português brasileiro, foi elaborada uma versão pelo setor de Lingüística Aplicada da PUCSP (Aluisio, Pinheiro e Finger, 2003).

[20] Constituído por um conjunto de textos, randomicamente selecionado, do jornal Folha de São Paulo, do ano de 1994, sendo composto por 1.1 milhão de *tokens*(Aluisio, Pinheiro e Finger, 2003).

sentença de TREINAMENTO é realizada a etiquetagem individual por cada membro do comitê de etiquetadores, permitindo a seleção de sentenças, com maior representatividade no aumento da exatidão dos etiquetadores, objetivando a correção manual. O fluxo de trabalho, que é concluído num intervalo de 24h, é ilustrado na Figura 27.

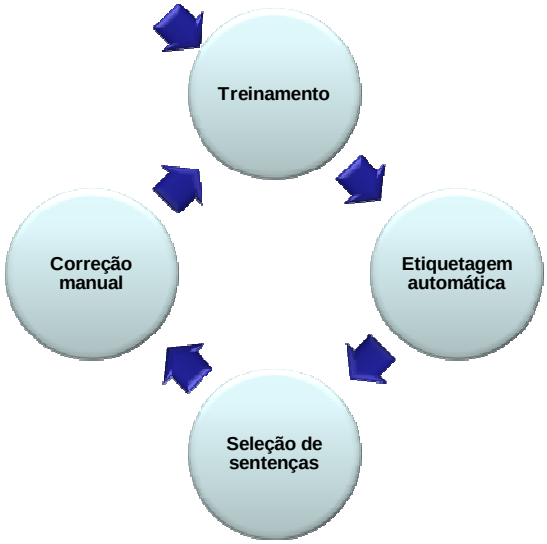


Figura 27 — Fluxo de trabalho com o *Active Learning*

A seleção de sentenças é realizada por meio do cálculo do índice de discordância de sentença $D_{sent}(s)$, função do índice de discordância de token $D_{tok}(t)$, para cada sentença sem marcação manual em uma determinada iteração. Tais índices expressam, em uma escala variável de 0 a 1, o grau de inconsistência de resultados obtidos pelo comitê de etiquetadores, ora em nível de *token*, ora em nível de sentença. As equações "Equação 9" e "Equação 10" (Tomanek, Wermter e Hahn, 2007), expressam matematicamente esse propósito. Nelas, $\frac{V(l_i,t)}{k}$ é a razão de k etiquetadores que atribuíram a etiqueta l_i para um token t e $|s|$ é o tamanho da sentença sob análise.

$$D_{tok}(t) := -\frac{1}{\log k}\sum_{l_i}\frac{V(l_i,t)}{k}\log\frac{V(l_i,t)}{k} \qquad \textbf{Equação 9}$$

$$D_{sent}(s) = \sum_{j=1}^{|s|}\frac{D_{tok}(t_j)}{|s|} \qquad \textbf{Equação 10}$$

A cada iteração de treinamento, o modelo criado para a *OpenNLP* é avaliado quanto à exatidão sobre o conjunto VALIDAÇÃO. O valor de incerteza encontrado em cada iteração e a média dos índices de discordância $D_{sent}(s)$ das sentenças sem marcação manual são armazenados para posterior análise gráfica de suas variações em função do número de *tokens* corrigidos a cada passo.

A etapa de correção manual é realizada por especialistas de domínio, com apoio da ferramenta desenvolvida especificamente para este propósito (Figura 28). Na ferramenta em questão, é apresentado ao especialista um conjunto de sentenças para correção, em que a etiqueta predominante (mais freqüente dentre as sugeridas pelo comitê de etiquetadores) para cada *token* é destacada e sugerida como correta, permitindo a correção manual no caso de indicação errônea. A ferramenta ainda exibe de forma gráfica os valores de exatidão e a média dos índices de discordância obtidos nas iterações antecedentes, além de destacar o último valor obtido.

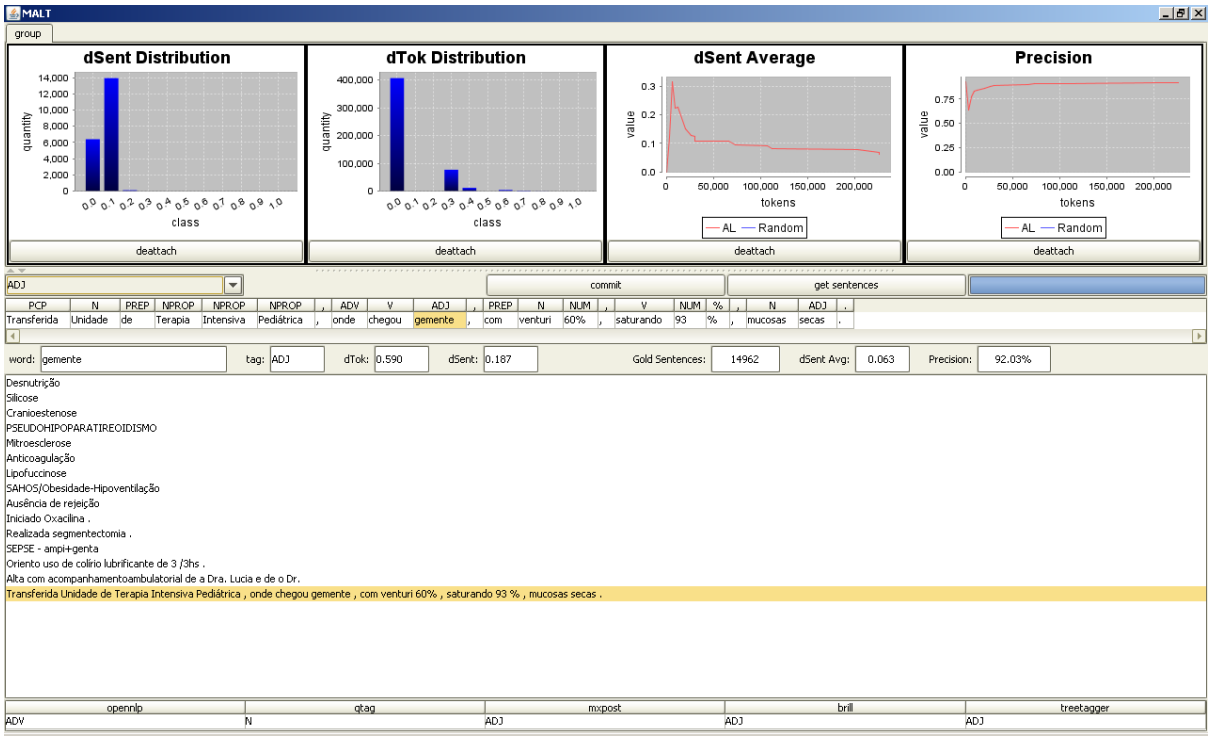


Figura 28 — Interface gráfica da ferramenta de correção manual de etiquetas

O ciclo foi interrompido quando do alcance de um índice estacionário de exatidão[21], após 160 iterações completas. A precisão alcançada é de 93.67%, compatível com iniciativas similares para outros idiomas (Hahn e Wermter, 2004).

[21] Indica que o índice de exatidão não sofreu variação após novas iterações.

4.3 Identificação de candidatos a frases nominais

Segundo Carone (2001), uma frase nominal é “constituída de qualquer dos elementos secundários que, na oração, se organizam em torno do verbo”, servindo para retratar algo de maneira estática.

A identificação de sentenças nominais é, normalmente, realizada com base em modelos construídos a partir de modelos de treinamento. A estratégia comumente utilizada remete aos modelos de máxima entropia.

Considerando o objetivo do trabalho, de mapeamento de texto livre para uma ontologia de domínio, e a premissa que os conceitos ontológicos são descritos de forma nominativa, foi desenvolvido um modelo de treinamento com base nas descrições de conceitos da SNOMED em espanhol, visto que não há tradução disponível para o português(IHTSDO, 2009).

A escolha ocorreu pela similaridade entre o português e espanhol, bem como o conjunto de evidências providas por (Schulz, Sbrissia e Nohama, 2004), que propuseram 45 regras de identificação cruzada, como os exemplos listados na Tabela 8, entre o português e espanhol, com bons resultados no domínio de Recuperação de Informação.

Tabela 8 — Exemplos de regras de substituição entre o português e o espanhol

Regra (P→S)	Exemplo em Português	Exemplo em Espanhol
qua→cua	Quadr	Cuadr
eia→ena	Veia	Vena
ss→s	Fracass	Fracas
lh→j	Mulher	Mujer
lh→ll	Detalh	Detall
l→ll	Lev	Llev
i→y	Enai	Ensay
f→h	Formig	Hormig
+ca→za	Cabeca	Cabeza
+o+→ue	Sort	Suert

Complementarmente, realizou-se o levantamento dos *tokens* mais freqüentes no *corpus* analisado e não presentes no dicionário de 8.921 *tokens* levantados para a correção ortográfica. Com base nesses resultados, um lingüista propôs a substituição de palavras e

morfemas funcionais espanhóis (conjunções, preposições, artigos, sufixos) para o português, como em: dos, de la, del, de los, de las, no, un, una.

As 325.687 descrições foram, então, etiquetadas, gerando uma lista de produções como a ilustrada na Tabela 9.

Tabela 9 — Amostra de produções etiquetadas com base na SNOMED em espanhol modificado, contendo o termo e a etiqueta gramatical identificada (ex.: o termo pedículo foi etiquetado com NADJ)

pedículo_NADJ de_PREPDE a_ART séptima_NADJ vértebra_NADJ cervical_NADJ
límite_NADJ de_PREPDE o_ART conducto_NADJ hialoide_NADJ
atención_NADJ de_PREPDE implante_NADJ auditivo_NADJ de_PREPDE o_ART ouvido_NADJ médio_NADJ
folículo_NADJ piloso_NADJ
corpo_NADJ de_PREPDE a_ART sexta_NADJ vértebra_NADJ dorsal_NADJ

O modelo de frases nominais foi, então, obtido com base nas freqüências dos padrões das etiquetas, conforme ilustrado na Tabela 10.

Tabela 10 — Amostra de padrões utilizados para a identificação de frases nominais obtidas a partir da SNOMED em espanhol modificado

Padrão de etiquetas	**Ocorrências**
NADJ NADJ	53725
NADJ	24843
NADJ NADJ NADJ	19922
NADJ PREPDE NADJ	18191
NADJ PREPDE NADJ NADJ	13462
NADJ PREPDE ART NADJ NADJ	10856
NADJ PREPDE ART NADJ	8392
NADJ NADJ PREPDE ART NADJ	6726
NADJ NADJ PREPDE NADJ	5721
NADJ NADJ PREPDE ART NADJ NADJ	5424
NADJ PREPDE NADJ NADJ NADJ	4453
NADJ NADJ NADJ NADJ	4199
NADJ NADJ PREPDE NADJ NADJ	4018
NADJ PREPDE ART NADJ NADJ NADJ	3289
NADJ PREPPOR NADJ	2797

Com base nos padrões identificados, um conjunto de expressões regulares foi criado, de forma automática, objetivando a obtenção dos candidatos a frases nominais. Para o propósito em questão, não foi realizada a desambiguação, neste sentido, no caso de uma

sentença formada por 6 palavras, $\{w_1, w_2, w_3, w_4, w_5, w_6\}$, é possível a identificação de candidatos com áreas de sobreposição como, por exemplo, $\{\{w_1\}_{f_1}, \{w_1, w_2\}_{f_2}, \{w_1, w_2, w_3\}_{f_3}, \ldots\}$; neste caso, o algoritmo apenas identificará as possíveis frases nominais. Para uso futuro, para cada candidato identificado, a freqüência do padrão utilizada é computada.

A validação do modelo adotado ocorreu com base na identificação manual de frases nominais no conjunto VALIDAÇÃO. Com base no levantamento realizado, foi considerado acerto no caso da frase nominal apontada pelos especialistas fosse uma das duas mais freqüentes do conjunto identificado pelo processo automatizado.

Com base nas premissas pontuadas, a precisão identificada foi de 91,03%, resultado compatível com estudos similares aplicado a outros idiomas (Voutilainen, 1995; Huang e Lowe, 2005; Bashyam e Taira, 2007).

4.4 Conclusões

A análise dos fenômenos lingüísticos do *corpus* análise permitiu identificar que erros de gramática (estilo telegráfico) são comuns e podem até ser classificados como fenômeno para-gramático característico do gênero de textos (caracterizando uma sublinguagem).

Para a grande maioria dos erros, a classificação utilizada engloba duas categorias: **sistemáticos e conscientes**, que são erros característicos de certos autores que parece terem desenvolvido um sistema pessoal otimizado para a produção de textos. A categoria dos erros sistemáticos inclui a omissão sistemática de pontuação, de acentos, ou de espaços. Enquadra-se nessa categoria também o fenômeno de capitalização notória e **lapsos involuntários**, que englobam os demais fenômenos. Porém, a ocorrência desse fenômeno revela um descuido voluntário com a qualidade do documento, pois esses erros poderiam ser facilmente corrigidos pelos autores.

Outra característica presente é a grande prevalência de acrônimos nas narrativas (100% do domínio manualmente analisado).

Estas características demandaram a adoção de uma solução de correção ortográfica, baseada na similaridade normalizada de Levenshtein, com eficácia de 86%, e de normalização, com base em um dicionário manualmente expandido.

Os modelos de etiquetagem morfológica foram construídos através da técnica de "*Active Learning*", utilizando a OpenNLP, com precisão de 93.67%. Esta precisão,

considerando as características dos documentos em análise (‘sujos’), superou a expectativa inicial do trabalho.

Com base na ferramenta de etiquetagem, foi possível construir o modelo de identificação de frases nominais, fundamental para a etapa de mapeamento. O modelo foi construído com base nas descrições nominativas dos conceitos da SNOMED em espanhol, sendo este funcionalmente adaptado para o português de forma automatizada.

Capítulo 5

5 NORMALIZAÇÃO MORFOSSEMÂNTICA

5.1 Introdução

Para o controle das variações lingüísticas, característica presente nos documentos analisados, adotou-se a normalização lingüística, categorizada em três eixos: morfológico, sintático e léxico-semântico (Arampatzis e Weide, 2000).

Na normalização morfológica, há redução dos itens lexicais através da fusão de candidatos que apresentem a mesma equivalência de classe conceitual. A estratégia de *stemming* objetiva a extração do radical de uma dada palavra (Porter, 2009), e a lematização, que reduz os adjetivos à forma masculina singular, os substantivos a forma singular, e os verbos na sua forma infinitiva, são os procedimentos mais conhecidos para a obtenção de candidatos a equivalências (Arampatzis e Weide, 2000).

Na normalização sintática, sentenças equivalentes são identificadas e conceitualmente equivalidas, como em “Ressecção de tumor do acústico pela fossa média” e “Ressecção pela fossa média de tumor do acústico”.

Na normalização léxico–semântica, as equivalências são determinadas quando do emprego de relacionamentos semânticos, como os da sinonímia (Jacquemin e Tzoukermann, 1999).

A tradicional normalização lingüística, como apresentada, não permite um eficiente controle de vocabulário para o domínio médico, especialmente para os propósitos de Recuperação de Informação, pois a terminologia médica é caracterizada por formas complexas de composição, derivação, e inflexão, tais como descrito por (Schulz e Hahn, 2006):

a) **“variação ortográfica**: diabetes mellitus, diabete mélito;
b) **derivação**: diabetes, diabético, diabéticas, antidiabéticos;
c) **composição**: *hiperprebetalipoproteinemia*;
d) **sinonímia**: nephro..., renal; estômago, gastr;
e) **abreviação**: AVC, ECG, DPOC;
f) **nomes próprios**: diclofenaco, Viagra, Parkinson, ...”.

Nesta perspectiva, a normalização[22] semântica dos documentos recuperados e dos critérios de consulta aplicados através da aplicação de um vocabulário em um determinado domínio, como o apresentado em (Schulz e Hahn, 2002), permitem uma melhor recuperação de informação. Para tanto, faz-se necessária a inclusão de um pré-processador baseado no vocabulário de interesse. Basicamente, quando do processamento, os documentos são normalizados pela atuação de um pré-processador, que realiza operações derivadas da normalização lingüística e conhecimento do domínio de interesse (no caso o médico Figura 29 - A). No momento da realização de uma consulta, os critérios inseridos são, igualmente, normalizados (Figura 29 - B), permitindo a recuperação das informações.

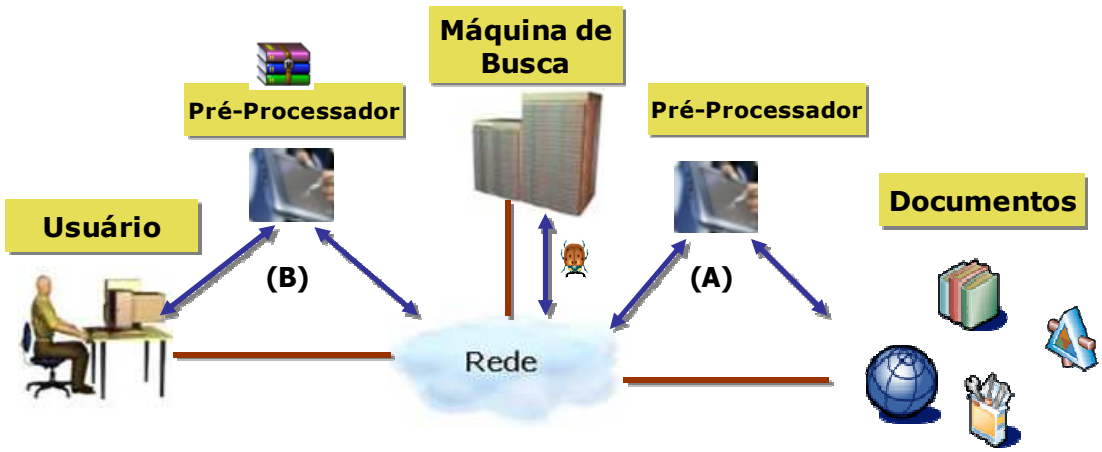


Figura 29 — Recuperação em máquinas de busca com pré-processamento baseado em um vocabulário de domínio

A arquitetura apresentada na Figura 29 melhora os resultados obtidos quando da recuperação de informação em domínios específicos, em especial, na área de saúde, em função das características da terminologia médica (Schulz e Mihov, 2002), por permitir o alargamento do universo de busca; no entanto, o cenário apresentado não considera elementos importantes que caracterizam as narrativas clínicas, tais como: uso de acrônimos, codificação

[22] Segundo o Dicionário Aurélio, normalizar é “**submeter a norma ou normas; padronizar”** (Ferreira, 1999).

de procedimentos clínicos baseados em diferentes tabelas, erros quando da escrita, entre outros.

Neste domínio, a escolha pelo sistema MorphoSaurus, como normalizador morfosemântico, enuncia estratégias para o processamento computacional adequado para narrativas clínicas. No entanto, o desenvolvimento do tesauro foi realizado ao longo dos anos especificamente para aplicação em contextos de Recuperação de Informação, no qual variações e incorreções menores produzem baixo impacto nos resultados obtidos.

No escopo do presente trabalho, no entanto, a qualidade do mapeamento está diretamente relacionada à qualidade do tesauro. Avaliações manuais iniciais indicaram a necessidade de readequação de estratégias e melhorias do tesauro para a utilização no escopo do presente trabalho.

5.2 Aspectos Sintáticos e Semânticos do tesauro

A construção do tesauro do MorphoSaurus é uma atividade complexa e manual, realizada por especialistas de domínio. Como qualquer atividade manual, a ocorrência de erros e decisões não ideais, são rotineiramente identificadas quando da análise do tesauro. Para o domínio de Recuperação de Informação, as incorreções afetam de maneira pouco significativa os resultados obtidos, dado a aplicação de estratégias de inter-relação entre sentenças em linguagem natural e as delimitações equivalentes formadas pelos MIDs (Markó, Daumke e Schulz, 2007). Para os casos mais específicos, como o mapeamento entre sentenças em linguagem natural e estruturas de domínio, as pequenas incorreções geram impactos significativos na qualidade dos resultados obtidos (Pacheco, Stenzhorn, *et al.*, 2009).

As formas gráficas (*tokens*) que constituem as palavras de um texto são muitas vezes ambíguas, podendo freqüentemente uma mesma forma corresponder a diferentes flexões de duas ou mais entradas lexicais distintas.

Como exemplo da complexidade decisória cita-se o caso típico da utilização de sufixos para a delimitação do radical, corr|er, corr|ido, etc. Porém, há numerosas exceções em que várias alternativas são concorrentes; como, por exemplo, reprodu|zir, reproduc|ao, reproduz|ir, na qual nota-se cada uma com duas variantes. Neste caso, procedimentos subjetivos são tolerados desde que as variantes do radical sejam igualadas de forma a preservar o conceito na geração da MID. O mesmo procedimento também é aplicado às variantes que, de certa forma, evitam erros de segmentação, como em nefro|tomia, nefr|o|tomia, nefro|otomia, onde podem coexistir sem dificuldade, pois seriam mapeados para o mesmo MID – desde que os conceitos sejam nefr=nefro e tomia= otomia.

Como se trata da construção de um tesauro com a inferência humana, é de se esperar que ocorram erros sistemáticos ou não. No decorrer da construção do tesauro, constataram-se erros tanto relacionados a um conceito ou classe, quanto também nos relacionamentos entre os mesmos.

Para o contexto do trabalho, foi necessário enunciar estratégias para a correção do tesauro, sem as quais o mapeamento de narrativas clínicas para uma estrutura ontológica mostrou-se inviável. Três foram as estratégias adotadas:

a) resolução de ramificações e ciclos,
b) detecção de idiossincrasias baseada na comparação de *corpora* similares (mesmo conteúdo, diferentes idiomas), e
c) validação de alterações baseada em *logs* do MorphoEditWeb.

Pelo fato de tratar de relações entre os conceitos, era comum haver no tesauro uma ramificação em uma direção desses relacionamentos, ou seja, um conceito #a relacionado para #b, o de #b para #c, de #c para #d e assim por diante. Normalmente, isso decorria de uma falta da delimitação precisa de um conceito, do domínio de um conceito. Em Markó, Schulz e Hahn (2003), constatou-se que esse tipo de configuração diminui o desempenho em sistemas de Recuperação de Informação. Alguns MIDS que apresentavam esta configuração são identificados na Tabela 11.

Tabela 11 — Lista de termos ramificados (cadeias)

hypoiiiqqia - brachijpjpya - lowgraderijxzia -
wayiijypya - tractiipzqra - railroadijxjwra -
extractiiiywra - tissueiijjjia - clothiiiwqka -
extractiiiywra - tissueiijjjia - clothiiiwqka - histoijixija -
fractureiiizzxa - intervaliijrppa - rhexisiipqjra –
baehniiipjq - wayiijypya - tractiipzqra - railroadijxjwra -
membraniikpzka - pessariikzwqa - phreniiyixia -
digitalisiiiwiza - dactylipxxwka - foxgloveriiqrxa -
controliijyyka - destinatiipqpxa - faehrtiiqypz -
ablatiiixixa - alleviatiikjyka - relievrijkqxa -
campijpywy - acreijqkyqa - arearikkyra - extensiiiwkxa -
campijpywy - acreijqkyqa - arearikkyra - extensiiiwkxa - domainrikkypa -
papuliikyxka - cysticercijqzzra - finnishriipwza –

Outro problema encontrado envolve o que se denomina de ciclos. Esse problema resume-se em fechar um ciclo que, partindo de um determinado MID, após mapear vários relacionamentos em série, voltar para esse mesmo MID. Ex.: uma classe #a que se relaciona com #b que se relaciona com #c que se relaciona com #a. Esse tipo de cenário impacta negativamente no processo de normalização do sistema MorphoSaurus. Alguns MIDs que apresentavam esta configuração, antes das correções, são identificados na Tabela 12.

Tabela 12 — Lista de termos com relações circulares identificados no Tesauro

baehniiipjq - **wayiijypya** - tractiipzqra - railroadijxjwra - **wayiijypya**
arrestikqwira - trapipxrxza - paradriipyx - terminiijqwpa - **verhaftriizpy** - festnehmriizqj - capturijqizia – **verhaftriizpy**
basaliiippqa - singliiiwzqa - easyiixppra – basaliiippqa
lowgraderijxzia - intensity_or_importance_or_widthrijyyia - widthiijirra - **intensityrijxyia** - importancrijxyja - size_or_intensity_or_importancerijxywa - extensiiiwkxa - groesseikqxqi – **intensityrijxyia**

Todos os 612 ciclos foram quebrados e as relações do tipo “cadeia” (1321) foram simplificadas, basicamente seguindo a regra (havia ramificações com até 8 nós):

Antes:

A has_sense B

B has_sense C

B has_sense D

Depois:

A has_sense C

A has_sense D

B has_sense C

B has_sense D

Considerando que o sistema MorphoSaurus é formado por um tesauro multilíngüe, existem casos de equivalência inconsistente, incompatível e incompleta quando da comparação entre idiomas. A validação intra e inter-língua do tesauro demanda o gerenciamento de equipes geograficamente separadas e que não falam o mesmo idioma (no caso do português x alemão x inglês).

Andrade (2007) evidencia diferenças na qualidade do tesauro multilíngüe do sistema MorphoSaurus, ao comparar os resultados da aplicação de Recuperação de Informação em *corpora* paralelos. Neste sentido, definiu-se uma estratégia de resolução de correlações inter-língua potencialmente incorretas.

A idéia, basicamente, consiste na comparação da freqüência dos MIDs em cada um dos idiomas considerados, quando da ocorrência de baixa correlação indica a necessidade da intervenção dos especialistas (Andrade, Pacheco, *et al.*, 2007). A operacionalização da atividade foi realizada no seguinte conjunto de ações:

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

Na preparação de *corpora* foram selecionadas fontes textuais nas línguas inglesa, alemã, portuguesa, espanhola e sueca (obtidas do *site* da Merck[23]). Seqüencialmente, os textos foram indexados pelo sistema *MorphoSaurus*, que permitiram a geração de tabelas de ocorrências para cada língua.

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

A abordagem aqui proposta está baseada na suposição que há uma correlação entre a freqüência de ocorrência das palavras no *corpus* em um idioma **A** comparado com a freqüência de ocorrência das traduções correspondentes no *corpus* comparável de um idioma **B** (Fung, 1998). É de se esperar que a distribuição de descritores semânticos (como as MIDS do MorphoSaurus) em cada *corpus* exiba um alto grau de similaridade. Na ocorrência de

[23] http://www.merck.com

grandes discrepâncias na distribuição de descritores semânticos (MIDs), tem-se um indício de problema no tesauro.

No processo de normalização de cada *corpus* estatístico, foram produzidas listas de freqüências das MIDs em cada *corpus* MSD, baseado na normalização morfossemântica do MorphoSaurus, permitindo a geração de listas com as freqüências de cada MID em cada idioma. O objetivo foi confrontar, de forma bilíngüe, as MIDs das listas e priorizar aquelas com maior discrepância de ocorrências. Para tanto, foi produzido um índice (*score* - S) que permite a classificação das MIDs concorrentes. Esse índice foi parametrizado de acordo com as equações 4, 5 e 6, de forma a assumirem valores entre 0 e 1; onde as MIDs, próximo da unidade, indicam uma maior probabilidade potencial de erro, demandando correção.

$$S = \frac{2S_d + S_a}{3} \qquad \text{Equação 11}$$

$$S_d = \frac{|f1 - f2|}{|f1 + f2|} \qquad \text{Equação 12}$$

$$S_a = \frac{fx}{(fx1 + fx2)_{\max}} \qquad \text{Equação 13}$$

onde:

- $f1$ é a freqüência da ocorrência de uma MID num *corpus*;
- $f2$ é a freqüência da ocorrência de uma MID em outro *corpus*;
- fx refere- se aos índices de cada linha de lista de MIDs comparáveis (de uma língua em relação às outras);
- $(fx1 + fx2)_{\max}$ corresponde a valor máximo da ocorrência do descritor em cada idioma;
- S_d expressa um índice com base na diferença de ocorrência de uma MID em um *corpus* normalizado em relação a outro;
- S_a 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 de problemas no tesauro normalizado entre 0 e 1.

Listas de freqüência foram geradas comparando pares de idiomas (como ilustrado nas Tabela 13 e Tabela 14), permitindo a identificação de potencias erros que exigem correção.

Tabela 13 — Amostra de freqüências das MIDs e seus parâmetros relacionados entre português ($f1$) e inglês ($f2$)

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

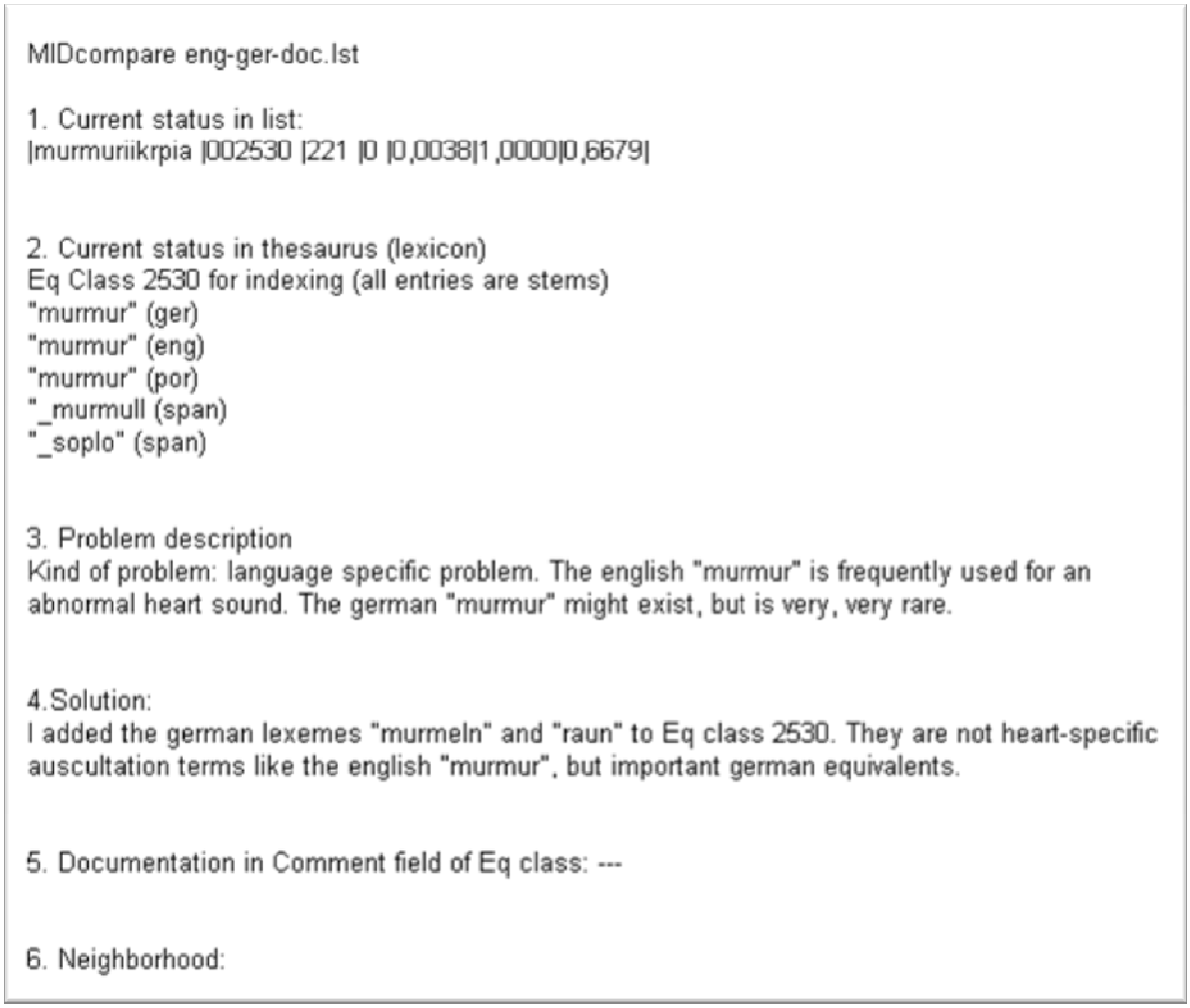
Tabela 14 — Amostra de freqüências das MIDs e seus parâmetros relacionados entre alemão ($f1$) e inglês ($f2$)

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

A inter-relação das equipes de lexicógrafos foi realizada através de um formulário padrão, elaborado em inglês, apresentado na Figura 30.

Para a determinação do padrão ouro, objetivando a avaliação da metodologia, foi utilizada a coleção OHSUMED (Hersh e Buckley, 1994), que é um conjunto de 348.566 documentos médicos clínicos extraídos do MEDLINE (de um total de mais de 7 milhões de documentos) que cobre todas as referências dos 270 jornais de cinco anos (1987-1991), bem como um conjunto de 106 consultas textuais escritas em inglês, cujo conjunto ideal de respostas, julgamento de relevância, foram identificadas por especialistas em saúde. Existe

um total de 16.140 pares de *queries* e documentos relacionados pelo julgamento de relevância (Hersh e Buckley, 1994).


MIDcompare eng-ger-doc.lst

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

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

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

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

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

6. Neighborhood:

Figura 30 — Protocolo de comunicação entre lexicógrafos – inglês e alemão

Concluídas as rodadas de avaliação e correções (aproximadamente 100 iterações) tendo como base as listas de freqüência comparativa entre pares de idioma, foram levantadas as curvas de precisão e revocação antes e após as correções, utilizadas, para tanto, o OHSUMED. A cobertura das bases idiomáticas sofreu incremento aproximado de 50% no espanhol, 7% no português, 31% no sueco, 3,6% no alemão e 4% no inglês.

Outra importante dimensão geradora de erros era oriunda das modificações concorrentes e distribuída realizada pela equipe de especialistas de domínio que trabalham na manutenção da base do tesauro, demandando a definição de validações de alterações baseado em *logs*.

Muitas vezes, pela complexidade inerente à terminologia representada, modificações são realizadas de forma ‘circular’ (especialista A cria uma relação, especialista B remove a relação anteriormente criada,...), por diferentes usuários, denotando a falta de consenso sobre a forma de representação em um determinado vocábulo.

Para a validação de alterações baseadas em *logs* (Bitencourt, Pacheco, *et al.*, 2007; Bitencourt, 2007), foi criado um registro de procedimento para identificar os processos anômalos ocorridos no processo de engenharia do léxico, utilizando o MorphoEditWeb.

A primeira anomalia a ser identificada é a de relacionamento, que denotam os cenários em que relações são ‘circularmente’ removidas e novamente inseridas como no caso apresentado na Figura 31.

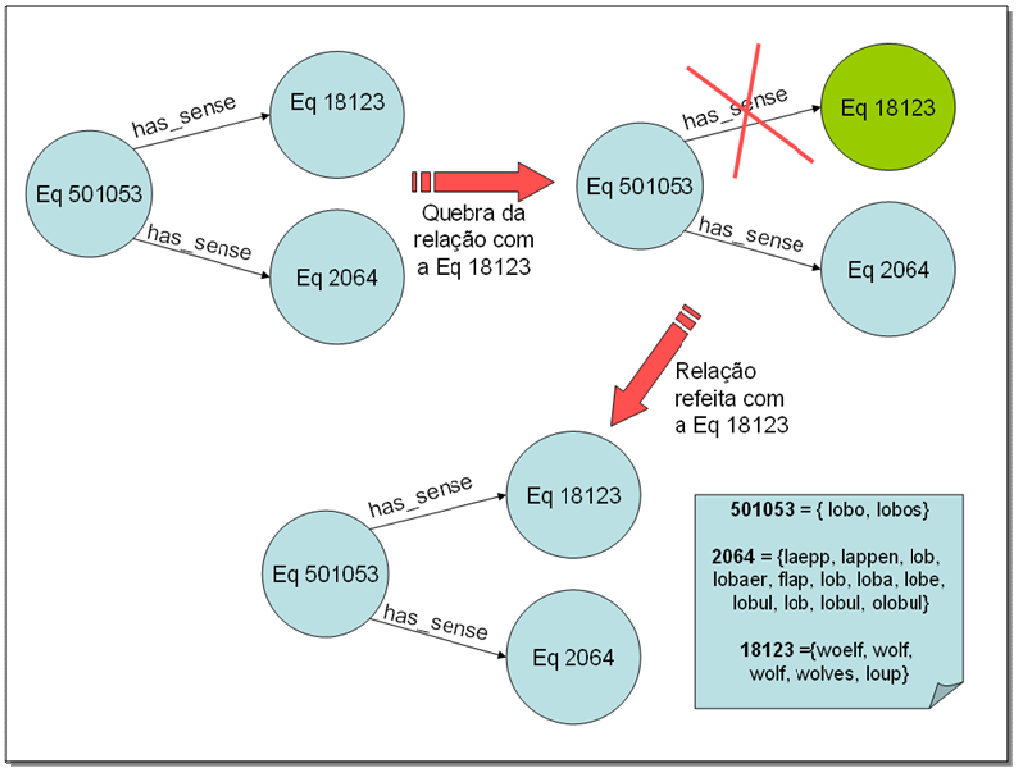


Figura 31 — Exemplo de anomalia de relacionamento (Bitencourt, 2007)

A segunda anomalia a ser detectada é a de tipo, decorrente das modificações em um lexema em uma determinada classe de equivalência. Os lexemas são agrupados em uma mesma classe para representar cenários de sinonímia. Neste caso, os especialistas inserem e removem um lexema, em uma classe de equivalência, sucessivas vezes, ao longo do tempo, como no exemplo apresentado na Figura 32, denotando a falta de consenso.

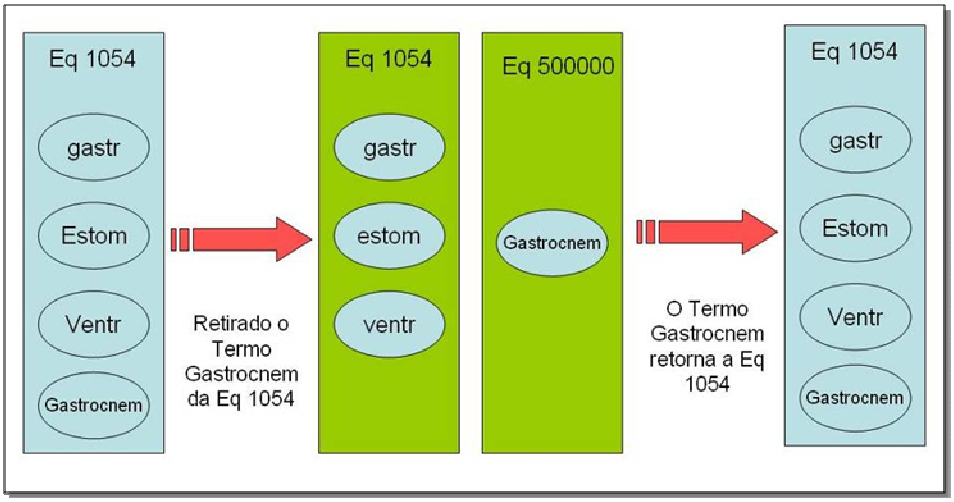


Figura 32 — Exemplo de anomalia de tipo (Bitencourt, 2007)

O trabalho da equipe de especialistas de domínio envolve a identificação, delimitação e cadastro de *subwords* (radicais, prefixos, sufixo e infixos), assim como as relações semânticas entre elas. Neste cenário, a anomalia de delimitação decorre de sucessivas modificações quando da delimitação de uma *subword* como no exemplo apresentado na Figura 33, normalmente decorrente da resolução de problemas de segmentação.

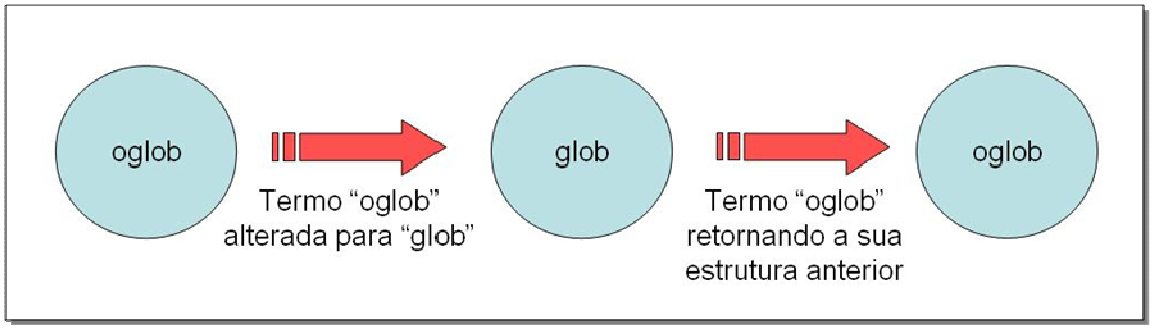


Figura 33 — Exemplo de anomalia de delimitação

A última anomalia identificada é a de permanência, quando uma *subword* é inserida na base, após sucessivas remoções.

As anomalias identificadas são regularmente enviadas para a equipe de administradores, objetivando a tomada de decisão em casos polêmicos, auxiliando no refinamento sucessivo do léxico do Sistema MorphoSaurus.

5.3 Conclusões

A terminologia médica é caracterizada por formas complexas de composição, derivação e inflexão, assim como pela geração contínua de novos acrônimos, abreviações e nomes próprios. Neste contexto, a indexação semântica aplicada aos documentos médicos, representa uma fronteira viável para a recuperação de documentos mais precisa e abrangente no domínio médico.

O sistema MorphoSaurus apresenta-se, metodológica e tecnologicamente, como estratégia fundamental para o tratamento computacional das complexidades inerentes à terminologia médica.

As ações realizadas com o intuito de refinar o tesauro produziram uma base com melhor qualidade e estabilidade, assim como definiram linhas de ações para a gestão do processo de engenharia de corpora.

Para detecção de idiossincrasias, baseada na comparação de *corpora* similares (Andrade, Pacheco, *et al.*, 2007), a atenção inicial focou-se no número de entradas do tesauro, em especial nas classes de equivalências e a presença de *subwords* sinônimas. Identificados os termos com maior grau de discrepância quando comparados em pares de idioma com textos paralelos, o foco voltou-se para a correção de segmentações errôneas, normalmente ligadas às questões sintáticas ou à criação de novas classes de equivalências.

Ao final das correções, verificou-se que bases com menor cobertura, como o espanhol e sueco, tiveram um incremento muito significativo na precisão (*benchmark* realizado com uma amostra de 106 *queries* do OHSUMED). Cabe destacar que a correção e/ou aumento da cobertura dessas bases foi guiada pelas listas geradas, permitindo uma administração eficaz das equipes internacionais de especialistas de domínio que efetuam o cadastramento terminológico.

As anomalias de relacionamento, parte da validação de alterações baseado em *logs* (Bitencourt, Pacheco, *et al.*, 2007), envolvem a análise e resolução de ambigüidades, resultantes de falta de precisão semântica, tornando o tesauro inadequado para o seu fim (Soergel, 1997). Nos 146 casos de anomalias identificadas, 41,09% foram recorrentemente debatidas entre os especialistas de domínio, nas listas de discussão. Vinte e três foram as classes que apresentaram essas anomalias.

Capítulo 6

6 MORPHOMAP

6.1 Introdução

No contexto médico, é reconhecida a necessidade de representar, explicitar e recuperar conhecimento, em decorrência da direta relação com a manutenção da vida e do bem estar de seres humanos. Neste sentido, inúmeros trabalhos estão relacionados à recuperação de informações na área médica (Honeck, U. e Klar R., 2002; Walczak, 2003; Parry, 2004; Kagolovsky e Moehr, 2004), normalmente através da indexação automatizada ou semi-automatizada de documentos. No entanto, comumente, são trabalhos centrados na localização de informações e não na valoração da informação recuperada no domínio considerado.

A capilarização dos conhecimentos médicos é uma área estratégica (Pan, Han e Yin, 2008), destacando-se a necessidade do uso de técnicas e metodologias que aliem os conceitos de recuperação de informação e mapeamento de conhecimento tácito e explícito, que, normalmente, seguem a estrutura definida no ciclo de Grundstein (1992), ilustrado na Figura 34 e, atualmente, representado através de estruturas ontológicas.

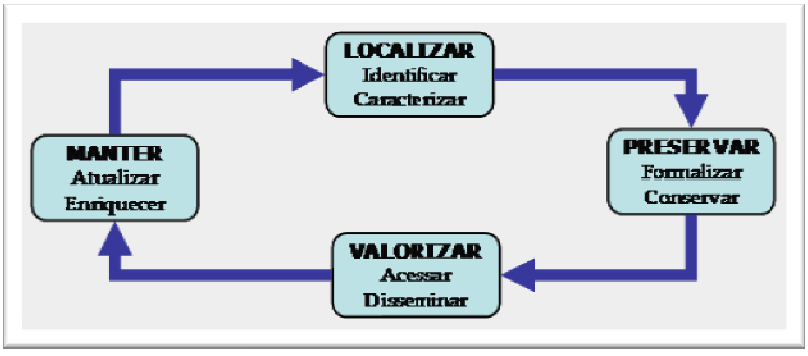


Figura 34 — Mapeamento de conhecimento segundo Grundstein (1992)

A aplicação de procedimentos baseados no ciclo de Grundstein permite que o conhecimento representado auxilie na execução de atividades e na tomada de decisão. Para a representação de conhecimento, dentre as estratégias disponíveis, apresentadas no Capítulo 2, o uso de ontologias apresenta-se como alternativa de maior aderência para o domínio médico, especialmente pela disponibilidade de artefatos com grande cobertura para a área médica, como a SNOMED CT (IHTSDO, 2009).

6.2 SNOMED CT

SNOMED *Clinical Terms*, SNOMED CT® (IHTSDO, 2009)[24] foi inicialmente desenvolvido como uma terminologia médica destinada a suportar a documentação em instituições de saúde de forma geral, para os diversos cenários cobertos pelo exercício clínico. SNOMED CT pode ser utilizado para codificar, recuperar e analisar dados clínicos.

SNOMED CT é resultado da união da SNOMED *Reference Terminology* (SNOMED RT), desenvolvido pelo colégio dos patologistas americanos, e o *Clinical Terms Version 3* - CTV3 - desenvolvido pelo *National Health Service* - NHS, Inglaterra.

A revisão constante da SNOMED CT, com a inclusão de relações ontológicas, permite classificar a SNOMED CT como uma ontologia em processo de revisão e concepção, ou ainda, sob uma linha conceitual menos estrita, como uma ontologia propriamente dita.

A SNOMED CT é composta por um conjunto de conceitos, termos e relações que objetivam a precisa representação de informação clínica, no domínio da saúde. A cobertura da terminologia é dividida em hierarquias (por exemplo: localização, drogas farmacêuticas, contexto social), objetivando facilitar a identificação do conhecimento representado (IHTSDO, 2009).

Os elementos básicos que compõem a SNOMED são (IHTSDO, 2009):

a) **conceitos**: unidade básica da SNOMED representa uma "unidade de significado". Um conceito é definido por um código numérico único, nome único (*Fully Specified Name*), um conjunto de termos (*descriptions*), um "*Preferred Term*" e sinônimos;
b) **hierarquias**: são 19 as principais, sendo que cada hierarquia admite sub-hierarquias;
c) **relações:** na SNOMED, as relações são do tipo "*is-a*", ou seja, objetivam a "ligação" entre conceitos; e
d) **descrições:** termos ou nomes atribuídos a um conceito.

[24] http://www.ihtsdo.org/

A SNOMED é distribuída sobre a forma de arquivos texto, *flat*, com quatro artefatos principais, relacionados segundo o modelo apresentado na Figura 35. Os arquivos são distribuídos separadamente, conforme o idioma. Cada distribuição (conjunto de arquivos que compõem a SNOMED) é disponibilizada de acordo com um modelo de licenciamento específico. As distribuições são qualificadas com o mês e ano.

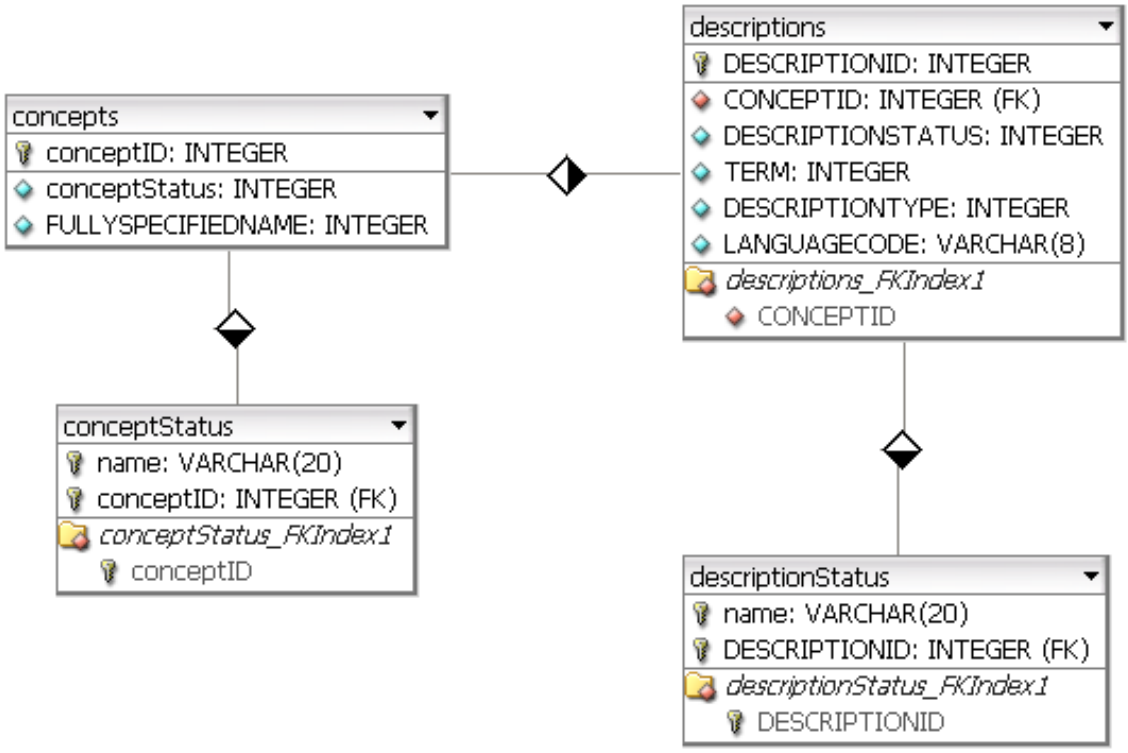


Figura 35 — Modelo relacional da distribuição da SNOMED CT

As tabelas ‘conceptStatus’ e ‘descriptionStatus’ são qualificadoras dos registros relacionados, seja de conceitos ou de descrições.

O conjunto de valores admitidos para ‘conceptStatus’, na distribuição considerada (janeiro/2009) são:

0. Corrente;
1. Retirado;
2. Duplicado;
3. Desatualizado;
4. Ambíguo;
5. Errado, e
6. Limitado.

O conjunto de valores admitidos para ‘descriptionStatus’, para a mesma distribuição, são:

0. Corrente;
1. Não corrente;
2. Duplicado;
3. Desatualizado;
4. Errado;
6. Limitado;
7. Inapropriado;
10. Movido para outra estrutura, e
11. Movimentação pendente.

O domínio para ‘*descriptionType*’, na tabela ‘*description*’ é o:

1. *Preferred Terms*: denota uma descrição padrão para o conceito relacionado*;*
2. *FullySpecified Name:* descrição composta por um dos *Prefered Terms* e a hierarquia de nível mais alto*;*
3. *Explanation or Definition*: definição, e
4. *External reference*: definição importada para o SNOMED CT.

O conjunto domínio para o campo ‘*LANGUAGECODE*’, da tabela ‘*description*’, para os idiomas espanhol e inglês, é:

a) en: descrições comuns para o idioma inglês;
b) en-GB: descrições específicas para o inglês britânico;
c) en-US: descrições específicas para o inglês americano, e
d) sp: descrições para o idioma espanhol.

A base para a exploração de um conceito é iniciada através da pesquisa pela tabela *concept*, com conceptStatus = 0 {*current*}. A tabela *concept* é independente de idioma. As descrições são obtidas pelo relacionamento de *concept* com a tabela *description*, através do *conceptId*. Na Figura 36, ilustra-se um exemplo da relação entre as tabelas *concept* e *description*, assim como os *status* associados, com base no conceito 84114007 (*Heart Failure*).

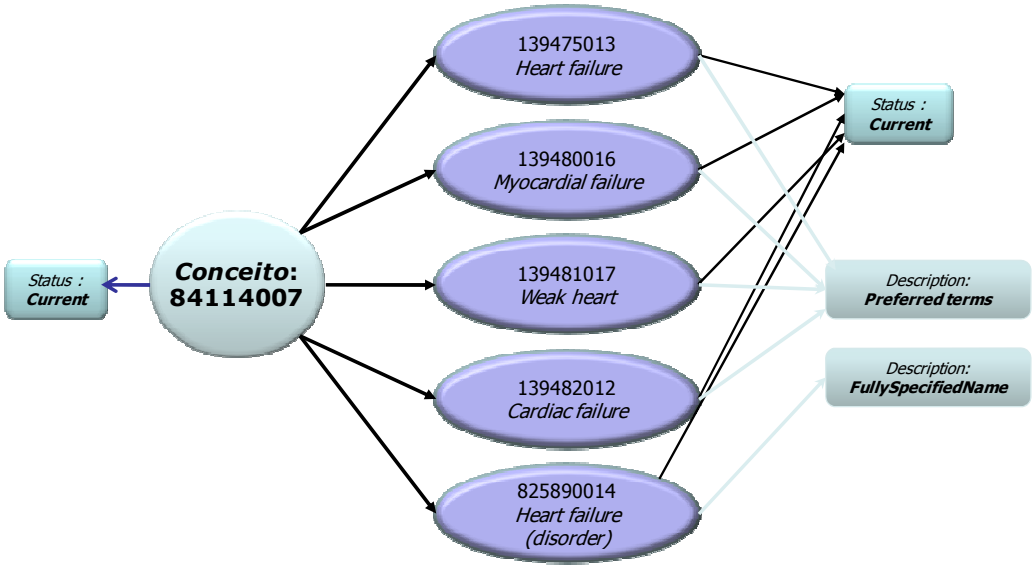


Figura 36 — Representação de instâncias para o conceito 84114007 e suas relações

6.3 Validação do padrão ouro para o mapeamento de texto livre para SNOMED CT

A validação do padrão ouro foi realizada com base nas 80 narrativas recuperadas do conjunto VALIDACAO (os prontuários selecionados foram submetidos à atividade de pré-processamento: correção ortográfica e expansão de acrônimos), que foram manualmente mapeados (para cada narrativa o esforço necessário médio para o mapeamento foi de 5h).

Para a avaliação da concordância do trabalho realizado, utilizou-se o teste Kappa (k) que é o procedimento estatístico utilizado para avaliar a confiabilidade de variáveis categóricas e nominais (Cohen, 1960). Kappa é interpretado como a proporção de concordância entre duas ou mais medidas de n observações, após a exclusão das concordâncias ao acaso.

Do total de documentos mapeados de forma única (por apenas um especialista de domínio), 12 documentos foram aleatoriamente selecionados e submetidos a todos os especialistas envolvidos na atividade de mapeamento. Com base na análise do mapeamento gerado, o Kappa obtido foi de 89%, indicando um elevado grau de concordância entre os especialistas envolvidos.

6.4 Mapeamento automático de linguagem natural para SNOMED CT

No processo de mapeamento de NPs para conceitos da SNOMED CT, o princípio multilingual do sistema MorphoSaurus é importante, visto que a SNOMED CT está disponível em inglês e espanhol mas não em português. Outro desafio presente no escopo do trabalho está relacionado com o tratamento de linguagem natural no que tange ao tratamento de variações lingüísticas específicas.

Para a operacionalização das atividades de mapeamento, foi constituída uma base derivada de uma distribuição da SNOMED CT, sendo aplicada, para cada conceito ativo, a normalização morfossemântica para as descrições associadas. O processo foi aplicado para descrições em espanhol e inglês.

Os MIDs resultantes foram obtidos com base na distribuição inglesa, conforme exemplo ilustrado na Tabela 15. Para os casos de ambigüidade (demarcadas com {}), nos quais um termo pode apresentar mais de um senso cadastrado na base do sistema MorphoSaurus, foram utilizados os termos não ambíguos, para o mesmo conceito, na base em espanhol.

Tabela 15 — MIDs para todas as descrições da SNOMED CT para o conceito "*Congestive heart failure (disorder)*"

SNOMED CT Concept Description	MIDs
Congestive heart failure	#abund #cardiac #deficien #static
Congestive heart disease	#abund #cardiac #disorder #static
Congestive cardiac failure	#abund #cardiac #deficien #static
CCF – Congestive cardiac failure	#abund #cardiac #ccf #deficien #static
CHF – Congestive heart failure	#abund #cardiac #chf #deficien #static

Cada ocorrência de um MID ambíguo é representada com um vetor binário em que cada posição indica a ocorrência ou ausência dos valores possíveis (ambigüidade). Um centróide é gerado para cada um dos sensos na etapa de treinamento. Os centróides são então comparados com os vetores representando os sinônimos para o mesmo conceito (em inglês e espanhol), para a determinação da similaridade. No caso de não resolução da ambigüidade com a estratégia apresentada, todos os MIDs presentes na cadeia ambígua são incluídos no resultado gerado.

Termos que não possuam equivalência para MIDs são mantidos na forma original, como na ocorrência de NOS, ilustrada na Tabela 16.

Tabela 16 — MIDs para o conceito 368009

Term	*MIDs*	*ID*
Heart valve disorder	cardiaciiirjk disorderiiiyii valvuliijxkz	1701013
Heart valve disorder, NOS	cardiaciiirjk disorderiiiyii nos valvuliijxkz	1702018
Valvular heart disease, NOS	cardiaciiirjk disorderiiiyii nos valvuliijxkz	1703011
Heart valve disease, NOS	cardiaciiirjk disorderiiiyii nos valvuliijxkz	1704017
Valvular heart disease	cardiaciiirjk disorderiiiyii valvuliijxkz	1705016
Heart valve disease	cardiaciiirjk disorderiiiyii valvuliijxkz	1706015
Disorder of heart valve	cardiaciiirjk disorderiiiyii valvuliijxkz	486555018
Heart valve disorder (disorder)	cardiaciiirjk disorderiiiyii disorderiiiyii valvuliijxkz	768595010

Considerando que determinadas hierarquias foram excluídas do processo de mapeamento, como "*product*" e "*context dependent categories*", conforme o *guideline*, foi constituída uma tabela com o conjunto de conceitos que devem ser ignorados no processo de mapeamento automatizado. O modelo de construção de representação morfossemântica baseada em MIDs (sistema MorphoSaurus) aplicado a SNOMED CT é ilustrada na Figura 37.

.

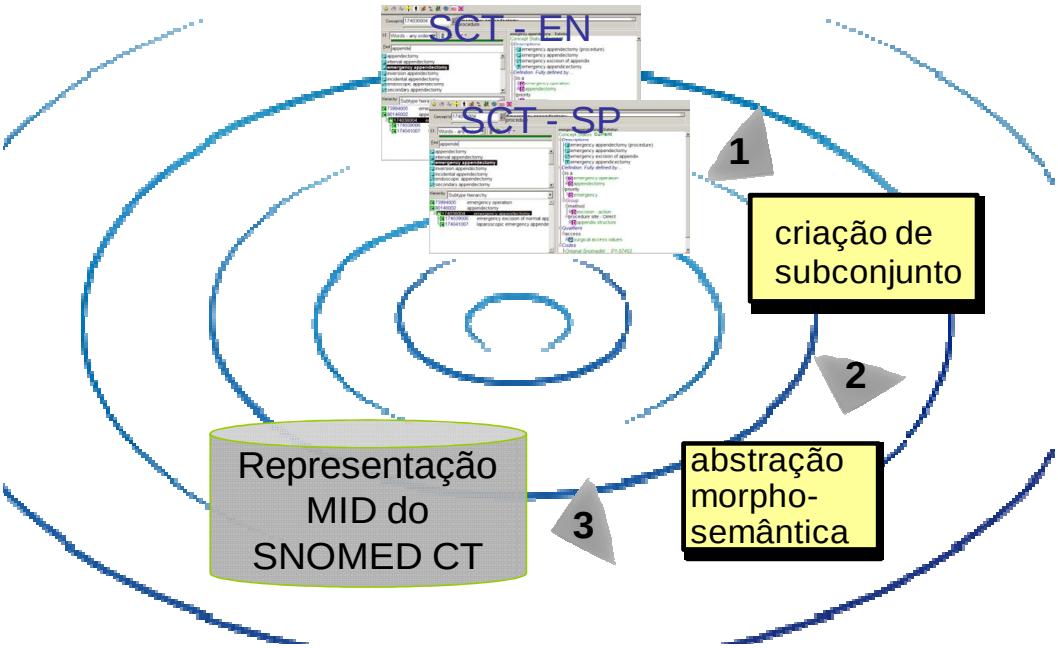


Figura 37 — Modelo de construção de representação morpho-semântica baseada em MIDs aplicado a SNOMED CT

Na Figura 38 ilustra-se o *pipeline* completo de mapeamento de linguagem natural (sumários de alta) para a SNOMED CT, que será detalhado na seqüência.

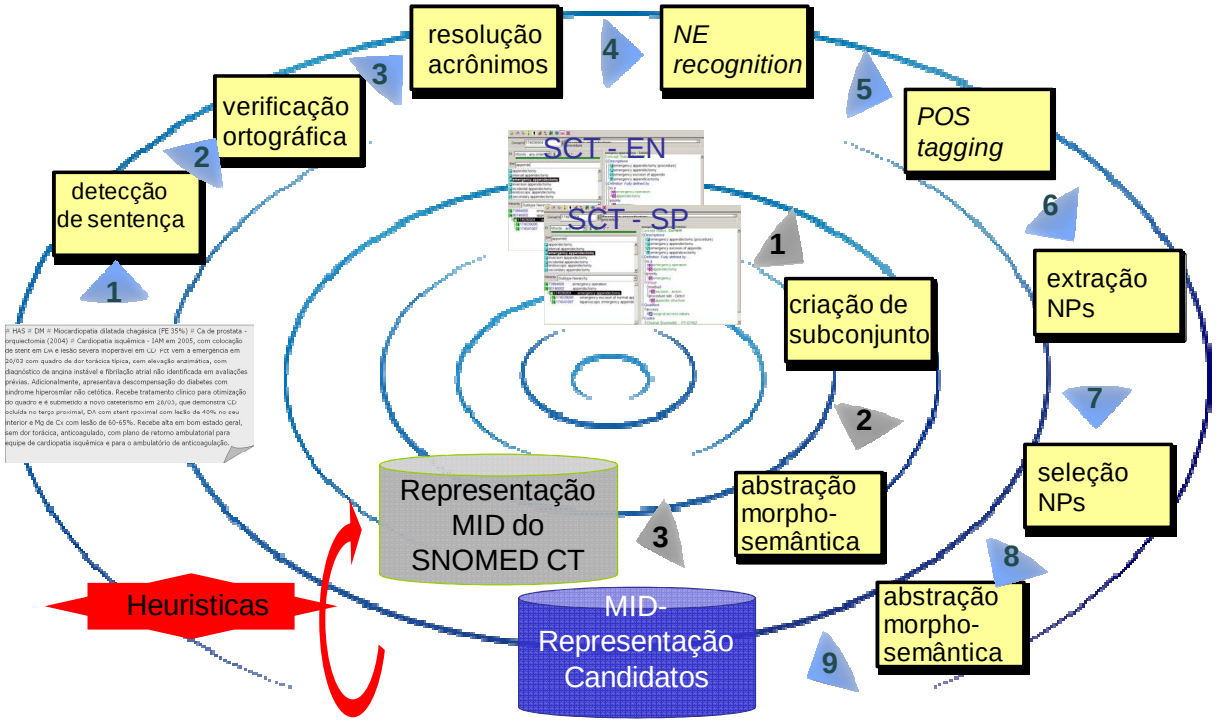


Figura 38 — *Pipeline* completo de mapeamento de linguagem natural para SNOMED CT

O processo de mapeamento é iniciado com a indicação de um sumário de alta. Com base no sumário, realiza-se a detecção de sentenças (normalmente delimitadas por fim de parágrafo ou ponto final).

Para cada sentença, de forma individual, aplica-se a correção ortográfica e a resolução de acrônimos, conforme estratégias detalhadas no Capítulo 4.

Na etapa subseqüente, realiza-se a identificação de nomes próprios. Entidades identificadas como nome próprio são marcadas para uso em aplicações futuras, especialmente para a identificação de novos nomes de drogas.

A etapa de *POS Tagging* consiste na categorização, para cada palavra, da etiqueta gramatical equivalente.

A extração de candidatos a frases nominais - NP, utilizada as etiquetas gramaticais anotadas na etapa anterior e os modelos constituídos com base nos SNOMED CT em espanhol, conforme apresentado no Capítulo 4. A identificação dos candidatos ocorre através de uma janela deslizante, de 2 até 5 palavras. A janela foi limitada em 5, dado que a ocorrência de sentenças com mais de 5 palavras é pouco freqüente nos sumários analisados (abaixo de 0,9%).

Para a seleção de NPs candidatos, são descartados os candidatos menos freqüentes, com base na técnica: descartar elementos que tenham ocorrência inferior a 40% do candidato

de maior ocorrência no conjunto de candidatos. O índice foi obtido com base em experimentos realizados objetivando a determinação da melhor relação entre qualidade e performance. Para cada candidato atribui-se um índice (chamado de NP_i). Para o elemento mais freqüente, NP_i é igual a 1. Para os demais elementos, aplica-se um decréscimo proporcional à posição do elemento considerado com base no conjunto de candidatos identificado (ex. para o cenário de 5 candidatos, o segundo elemento vale 0,8, o terceiro vale 0,64, o quarto vale 0,512 e o último vale 0,4096 – decréscimo de 20%).

A etapa de abstração morfossemântica consiste na normalização (obtenção dos MIDs) de cada uma das NPs extraídas. Como aproximação pode-se entender que os MIDs que compõem uma NP representam uma *query* de busca, já os conjuntos de descrições que compõem um conceito representam um documento. Os documentos são indexados com base em uma máquina de busca baseada no projeto Apache Lucene (APACHE, 2009), que permite a obtenção dos conceitos candidatos, para cada NP, bem como o respectivo tdidf normalizado. Documentos com menos de 60% dos MIDs considerados são descartados. Para os casos de MIDs que pertençam ao '*Prefered Term*' do conceito, o peso do termo na busca, para o cálculo do tdidf, do resultado é dobrado. O modelo das inter-relações é ilustrado na Figura 39.

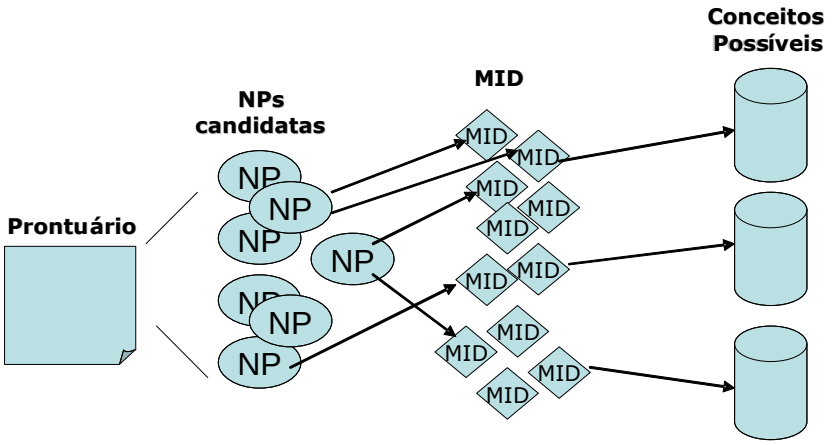


Figura 39 — Modelo de mapeamento para a etapa de construção morfossemântica

A seleção de conceitos passa pela determinação dos índices de mapeamento, determinados sob a forma: $NP_i * 0{,}10 + \text{tdidf} * 0{,}9$.

A seleção do primeiro conceito, com maior índice de mapeamento, gera a exclusão de possíveis conceitos, visto que a NP selecionada é sub ou super conjunto de outras NPs candidatas. Para o segundo nível de mapeamento, o índice é acrescido em 0,1 para os casos em que o par de conceitos (o mapeado na iteração anterior e o candidato) possuam relação

semântica conhecida (obtida através da aplicação de função de inferência entre os itens considerados).

Com base nos 80 documentos utilizados para a validação, foram identificados, automaticamente, 3.020 conceitos contra 2.901, manualmente identificados.

O resultado foi validado com duas estratégias: na primeira, para cada documento submetido ao processo de mapeamento, um dos especialistas que não teve acesso ao documento em qualquer uma das etapas anteriores, validou o resultado gerado; na segunda, a validação foi realizada com base no padrão ouro, sem margem à interpretação, desta forma, no caso de os conceitos mapeados pelo MorphoMap corresponderem ao mapeamento original considera-se um certo, caso contrário não.

Para a primeira estratégia, o índice de acerto foi de 83,9% (comparação entre o resultado manual e o automático). Para a segunda estratégia, o índice de acerto foi de 78,2%.

Subseqüentemente, as diferenças da segunda estratégia foram manualmente analisadas pelo comitê de especialistas e categorizadas (conforme resultados apresentados na Tabela 17).

Tabela 17 — Categorização das diferenças entre os resultados do mapeamento automático dos documentos do padrão ouro e o mapeamento automático realizado pelo MorphoMap

Categoria	%	Descrição
Extração de NP	20	NPs identificadas não corresponderam à pontuada pelo especialista, gerando mapeamento errôneo
Morpholization	24	Termos não mapeados para MID (~20% dos casos identificados, indicando necessidade de revisão do tesauro do sistema MorphoSaurus para estes casos), ou foram mapeados incorretamente (erro de relacionamento no tesauro) ou a ambigüidade não foi resolvida (~40% dos casos analisados)
Contexto	36	Cenários em que o especialista identificou dois ou mais conceitos, enquanto que o mapeamento automático selecionou apenas um, na mesma direção existem casos em que a negação não foi corretamente considerada (os casos de negação corresponderam a 60% dos casos nesta categoria)
Erros não corrigidos pelo processo de correção ortográfica	4	Situações em que a correção ortográfica não resolveu o erro de ortografia
Acrônimos não expandidos	1	Cenários em que não houve expansão de acrônimos
Outros erros	12	Situações de mapeamento que não foram categorizadas nas categorias anteriores

6.5 Detecção de subespecificação na estrutura ontológica da SNOMED CT

O processo de garantia de qualidade e auditoria desempenha um importante papel na manutenção de grandes terminologias e/ou ontologias, como a SNOMED CT. Inúmeras técnicas automatizadas têm sido propostas para facilitar a identificação de pontos frágeis e a proposição de melhorias (Pacheco, Stenzhorn, *et al.*, 2009).

SNOMED CT é uma grande base terminológica clínica, em função de um conjunto de fatores, conforme apresentados em (Pacheco, Stenzhorn, *et al.*, 2009):

a) a SNOMED CT foi constituído da junção de dois sistemas legados (SNOMED RT e NHS *Clinical Terms Version* 3), formados por princípios arquiteturais muitas vezes contraditórios;
b) em face das constantes demandas para a inclusão de conteúdo, bem como a origem histórica, a avaliação de pertinência nem sempre foi criteriosa; e
c) os processos de manutenção e auditoria não acompanharam o crescimento da base.

Considerando a quantidade de conceitos na SNOMED CT, é improvável que os processos de manutenção, auditoria e garantia de qualidade sejam realizadas de forma completamente manual. Inúmeros métodos semi-automatizados têm sido propostos para identificação de problemas no conteúdo e na arquitetura (Bodenreider, Smith, *et al.*, 2007 ; Jiang e Chute, 2009; Wang, Halper, *et al.*, 2007).

Um problema conhecido na estrutura da SNOMED CT é a subespecificação (conhecido por *underspecification*) (Jiang e Chute, 2009). Em contraste aos erros “reais”, tem-se a falta de elementos e não a presença de elementos errôneos. A vantagem deste tipo de situação é que é a possibilidade de remediá-la de forma monótona (manual), sem a remoção de conteúdo (que gera impacto visto que os conceitos podem já ter sido utilizados). O cenário apresentado pode ser identificado em diversos conceitos da SNOMED CT que, apesar de suas descrições apresentarem significados compostos, não estão logicamente relacionados a nenhum outro conceito, a não ser seus parentes taxonômicos.

Observando a estrutura hierárquica na SNOMED CT para o conceito “*Cerebral function”*, observa-se que o mesmo está relacionado somente ao pai “*Nervous system function”*, em contrapartida, a relação esperada com o conceito “*Brain structure”* está ausente. Porém, como o conceito “*cerebral*” é derivado de “*cerebrum*” (como um sinônimo

de "*brain*"), uma metodologia baseada no uso de um léxico poderia inferir a relação lógica ausente. A inclusão desta metodologia no processo de construção da SNOMED CT pode auxiliar no processo de construção, manutenção e melhoria contínua da estrutura do tesauro.

Devido à grande diversidade de expressões na linguagem natural em termos de inflexões, derivações e sinonímia, realizou-se uma abstração conceitual dos significados de cada descritor. Precisamente, mapeou-se uma seqüência de *tokens* textuais (t_1, t_2, t_3,…,t_m) para uma seqüência de identificadores semânticos (m_1, m_2, m_3,…,m_n), utilizando o Sistema MorphoSaurus.

A indexação morfossemântica foi realizada para cada um dos 837.105 termos ativos da base Descrição da SNOMED CT, perfazendo uma média de 4,95 MIDs por descrição.

Para a seleção de Candidatos Subespecificados, foram selecionados conceitos sem relacionamentos seguindo os seguintes critérios: primeiramente, utilização dos conceitos ativos somente; em segundo lugar, exclusão de todos os conceitos que possuíam relacionamentos além de *is-a* (relação de classificação taxonômica).

Os atributos de alguns conceitos da SNOMED CT são relacionados a pares de conceitos através da tabela de relacionamento. Por exemplo, o conceito *Inflammatory disease of liver* tem como atributo *Finding site*: *Liver structure.*

Em contraste, o conceito *Hepatitis notification* não possui nenhum atributo, embora houvesse a expectativa de possuir um relacionamento com o conceito *Inflammatory disease of liver*.

Neste cenário, busca-se a identificação de atributos ausentes. Para o escopo da presente tese, optou-se por ignorarem-se os relacionamentos nativos, com foco exclusivo na estrutura de conceitos. A razão para esta decisão é que freqüentemente não era muito claro qual relacionamento existente utilizar ou ainda se seria necessário introduzir um novo relacionamento na SNOMED CT.

Nesse processo, desenvolveu-se a seguinte abordagem:

Assumiu-se que C seja um conceito sem atributos, que FSN_C seja um nome completamente especificado e que $P_C = \{P_1(C), P_2(C), \ldots, P_k(C)\}$ seja um conjunto de conceitos pais num relacionamento direto. Desta forma, para cada relacionamento direto $P_i(C)$, um FSN é usado: $FSN_{Pi(C)} = FSN(P_i(C)$.

Com base nessas regras, comparou-se as seqüências de MIDs de cada elemento de FSN_C com a seqüência de MIDs de cada $FSN_{Pi(C)}$ como segue: cada MID que ocorria em ambas as seqüências era eliminada da seqüência da primeira. Para a seqüência de MIDs restante, foi verificado quais correspondiam exatamente a seqüência MID de qualquer outra

descrição, em todo o conjunto das descrições da SNOMED CT (aqui não somente FSNs). Nesse caso, o conceito pertencente àquela descrição era sugerida para refinar o conceito original.

Para a validação dos tipos semânticos (apresentado pelas expressões entre colchetes no FSN, e.g. *Organism, Substance, Body Structure*), uma amostra aleatória de vinte conceitos sem relacionamentos foram extraídos e listados junto com todos os atributos refinados candidatos ao sistema proposto. Para cada conceito da amostra com um domínio especializado foi verificado (i) quais dos conceitos poderiam ser refinados e (ii) quais dos candidatos sugeridos, realmente poderiam ser utilizados no refinamento.

No sentido de validar os resultados, um segundo domínio especializado foi utilizado para realizar as mesmas verificações com metade das amostras (10 conceitos em cada hierarquia).

Aproximadamente metade dos conceitos da SNOMED CT (45,5%) não possui atributos. Com a aplicação da metodologia identificaram-se 48.552 (16,6%) conceitos como refináveis, sugerindo uma média de 2,8 conceitos alvos em potencial (que juntamente com uma relação sugerida pela estratégia, seria possível apurar a descrição lógica do conceito sob análise).

A avaliação foi realizada por dois especialistas de domínio (com formação médica e participante do grupo de desenvolvimento da SNOMED CT).

A Tabela 18 fornece uma foto exata da classificação pela hierarquia principal da SNOMED CT. De acordo com a estimativa baseada na amostra analisada, aproximadamente 18.500 conceitos foram refinados e mais de 12.000 conceitos foram sugeridos pelo sistema como “corretos”.

Uma verificação na distribuição revela que em algumas hierarquias um conceito não simples, é apresentado freqüentemente como atributo. É o caso de *Organism*, *Substance*, *Qualifier Value*, *Observable Entity*, *Physical Object* e *Occupation.* Isso é consistente com as diretrizes da edição da SNOMED CT aplicadas até agora. Entretanto, a metodologia apresentada sugere refinamento de conceitos para as hierarquias, ex., *Macaroni* for *Macaroni maker* (occupation), *Canada for Salmonella canada* (organism), *Metal for Metal device* ou *Acyl carnitine for Acylcarnitine hydrolase (substance).*

Estrutura corporal é outro caso interessante, já que se rejeitaram todas as sugestões de conceitos candidatos fornecidos e aceitou-se apenas uma das vinte no julgamento para refinamento. A razão disso é modo idiossincrático de emular hierarquias tipo *part-of* por taxonomias nomeadas de conceito “estrutura” ou conceito “parte” segundo o modelo

atualmente adotado na SNOMED CT (*SEP Triplet* (Schulz, Romacker e Hahn, 1998). Pois, as relações *part-of* já se apresentavam no modelo (embora mascarado pela construção SEP): *Cardiac wall structure* ***is-a*** *Heart Part.* Nesse caso, os conceitos alvos propostos mostraram-se inúteis. Também rejeitou-se as sugestões de refinamentos para certas partes do corpo com números ordinais na nomenclatura como *Fifth metatarsal structure by Five.*

Tabela 18 — Análise de conceitos subespecificados na SNOMED CT organizado por sub-hierarquias. (Pacheco, Stenzhorn, *et al.*, 2009)

		Conceitos subespecificados		Candidados a refinamento		Análise de exemplos(n=20)		Sample based estimation	
Hierarquias SNOMED CT	Conceitos ativos	n	%	n	%	Refinamento justificado	Sugestão correta	Conceitos refináveis	Sugestões corretas
Organism	31840	31840	100,0	4973	15,6	0%	0%	0	0
Substance	23554	23554	100.0	8627	36,6	55%	35%	4700	3000
body structure	25637	22386	87.3	15076	58,8	5%	0%	800	0
qualifier value	8823	8823	100.0	3533	40,0	0%	0%	0	0
observable entity	7885	7885	100.0	3647	46,3	70%	50%	2600	1800
Finding	32780	5356	16.3	2253	6,9	90%	75%	2000	1700
physical object	4408	4408	100.0	1339	30,4	85%	80%	1100	1100
morphologic abnormality	4297	4289	99.8	2164	50,4	80%	60%	1700	1300
Occupation	3843	3843	100.0	1330	34,6	75%	10%	1000	100
Product	19310	3541	18.3	686	3,6	100%	60%	700	400
Event	3578	3529	98.6	447	12,5	85%	45%	400	200
Disorder	63874	2812	4.4	1080	1,7	90%	60%	1000	600
Procedure	47764	2256	4.7	1001	2,1	85%	65%	900	700
Others	14511	7603	52.4	2396	16,5	75%	60%	1800	1400
TOTAL	292104	132125	45.2	48552	16,6			18700	12300

Uma razão quase que comum para se rejeitar o sistema de classificação de um conceito como refinável é que o mesmo já se apresentava suficientemente definido pelos relacionamentos pai, como, por exemplo, *Female first cousin* pela intersecção com *First cousin* e *Female cousin.*

O índice Kappa entre os dois especialistas foi de 0,74.

6.6 Conclusões

Na análise dos resultados do MorphoMap apontou-se precisão de 83,9%, com a co-validação do mapeamento automático por um especialista, ou de 78,2%, com a validação entre os documentos manual e automaticamente mapeados.

Na “Detecção de subespecificação na estrutura ontológica da SNOMED CT”, apresentou-se a metodologia de apoio à auditoria de construção de ontologias, no cenário do experimento a mesma foi aplicada sobre a SNOMED CT de forma a apontar falhas na especificação de definições do conceito lógico através da exploração das descrições presentes. Pela comparação de uma representação semântica simplificada do significado dos nomes dos conceitos e seus relacionamentos, a abordagem gera hipóteses a respeito dos possíveis atributos ausentes. Com base em uma análise manual de amostras aleatórias, estimou-se que aproximadamente 18.000 conceitos da SNOMED CT podem ser refinados. A análise por dois especialistas apontou *kappa* de 0,74.

Capítulo 7

7 DISCUSSÃO E CONCLUSÕES

7.1 Discussão

Na presente tese, empregou-se uma abordagem mista para a resolução de acrônimos e correção ortográfica, decorrente das características específicas das narrativas clínicas. Diferentemente de trabalhos que partem da premissa da existência de amplos bancos de dados, com inúmeras entradas para a expansão de acrônimos e/ou a correção de termos (Pustejovsky, Cochran e Morrell, 2001; Wren e Chang, 2005; Schwartz e Hearst, 2003; Ringlstetter, Schulz e Mihov, 2007; Veroyatnostei e Statistika, 2005), ou ainda, como em Pustejovsky *et al.* (2001) e Wren e Chang (2005), que utilizaram o processo de *stemming* ou o MEDLINE com a normalização de termos como parte do processo (Pustejovsky, Cochran e Morrell, 2001; Wren e Chang, 2005; Okazaki e Ananiadou, 2006). Não houve possibilidade de consituição automatizada de dicionário de expansão, visto que não há evidências de escrita conjungada entre acrônimos e seu significado nas narrativas clínicas (diferentemente de textos científicos); no domíno da correção ortográfica, diferentemente dos trabalhos citados, a prevalência de erros ortográficos é elevada, possivelmente pelo cenário clínico característico (pressão de tempo pelos cenários de urgência e emergência).

Apesar das diferenças e dificuldades características, os resultados obtidos são compatíveis com os trabalhos citados, conforme resultados comparados apresentados por Wren e Chang (2005).

Após uma década de pesquisa intensiva, a recuperação de informação entre diferentes idiomas, também conhecida por *Cross-Language Information Retrieval* - CLIR, produziu consideráveis resultados (Gey e Kando, 2002). No aspecto metodológico, é possível dividir o

campo do CLIR em duas abordagens distintas: a baseada em um dicionário e a baseada em um corpus de domínio (Markó, Daumke e Schulz, 2007). A abordagem baseada em *corpus* é a que demanda maior quantidade de recursos (humanos e materiais) para a construção. Neste sentido, a maioria das pesquisas nesta área está direcionada à tradução de consultas e/ou dos documentos (Darren e Allen, 2003). Markó *et al.* (2003) destacam que a tradução tanto de consultas como de documentos apresenta melhor performance, quando comparada às estratégias de tradução apenas da consulta.

No domínio das terminologias médicas, bem como de outras sublinguagens, poucos são os léxicos multilíngües disponíveis (Schulz e Han, 2000). O sucesso de aplicações CLIR baseadas em dicionários depende diretamente da cobertura do léxico, bem como das ferramentas de manipulação do mesmo (Hedlund e Kalervo, 2001). Com o sistema MorphoSaurus, a boa cobertura do léxico é garantida com delimitação do escopo do léxico ao sub-domínio médico. Os resultados de recuperação trans e intra-lingual da ferramenta e do tesauro foram apresentados por Markó *et al.* (2007) e Pacheco *et al.* (2009).

O sistema MorphoSaurus, bem como o léxico que o compõe, tendo sido constituído como um projeto internacional, envolvendo principalmente pesquisadores do Brasil e Alemanha, permitiu a fundamentação multi-lingüística do projeto desde os seus primórdios. Apesar dos significativos resultados na área de IR, em especial na área de CLIR, as incorreções do tesauro, naturais pelos processos humanos envolvidos na sua constituição, demonstram impacto negativo quando da utilização da ferramenta em processos mais sensíveis, como o de mapeamento. Nesta dimensão, as propostas de “detecção de idiossincrasias baseada na comparação de *corpora* similares” e de auditoria, mostraram-se extremamente eficientes na identificação pontual de problemas, elevando o patamar de desempenho do tesauro com relativa rapidez e baixo investimento de recursos humanos. Somente com a adoção das estratégias pontuadas foi possível propor a utilização do sistema MorphoSaurus em casos reais e não a cenários controlados, como os de trabalhos anteriores (Markó, Schulz e Hahn, 2003), focados na análise de documentos científicos.

Recuperar conhecimento armazenado em narrativas clínicas é objeto de estudo de diversos trabalhos (CHAPMAN *et al.*, 2004; DOBRYNIN *et al.*, 2005; GHOSH e SCOTT, 2005; HOODA *et al.*, 2004; ORLIKOWSKI, 1992; RUCH *et al.*, 2006), especialmente pelo intenso investimento no século XX no sentido de estabelecer as bases conceituais e tecnológicas para o registro de eventos clínicos (PESTOTNIK, 2000), produzindo grandes bases de dados e conhecimento que, agora, exigem novas técnicas para a sua manipulação.

O mapeamento de artefatos para estruturas formais tem recebido atenção recente da comunicação científica, com a publicação de trabalhos relevantes na área (JOSEPH *et al.*, 2006; MAEDCHE e STAAB, 2001; QAMAR e RECTOR, 2006; RUCH *et al.*, 2006). Como elemento diferenciador, o presente trabalho trata do processo de recuperação e mapeamento contemplando todo o ciclo: do registro do conhecimento (texto livre), passando pelo mapeamento e recuperação da informação desejada. A originalidade também está associada ao trabalho com a língua portuguesa em um domínio real e o uso de uma ontologia de domínio médico, no caso a SNOMED CT, a ontologia médica de maior cobertura na área que, no cenário da presente pesquisa, é utilizada para representar, de maneira padronizada, asserções relatadas em narrativas clínicas.

No domínio de mapeamento de textos científicos para uma estrutura formal, em destaque o trabalho de Markó *et al.* (2003), que utilizou o sistema MorphoSaurus para realizar o mapeamento semi-automatizado para o MeSH. Nesse trabalho, os autores utilizaram um sistema de força bruta na busca dos melhores identificadores para cada documento analisado. A estratégia proposta não pode ser aproveitada no escopo do presente trabalho pois partia-se da premissa que um *abstract* deveria ter um conjunto correlato de conceitos associados, com reforço positivo para as intercorrências. No caso de documentos reais, cada sentença pode apresentar mapeamento diferenciado, talvez complementar, mas o reforço positivo não ocorre neste cenário.

Referente ao mapeamento com o SNOMED CT, Want *et al.* (2007) propuseram a análise estrutural da terminologia, dividindo-a na área de estrutura e de relações. A área de estrutura era composta dos elementos hierárquicos e estes foram utilizados na identificação de similaridade entre conceitos e na identificação de potenciais novas relações na estrutura do SNOMED CT. Na mesma direção, Jiang e Chute (2009) propõem o uso da análise formal de conceitos, permitindo a identificação de potenciais relações faltantes ou na avaliação de hipótese de relacionamento entre elementos externos e a estrutura conceitual do SNOMED CT. Ambas as abordagens utilizam-se da premissa que novos conceitos seriam codificados na forma pós-coordenada, metodologia descartada no início dos trabalhos pela inviabilidade prática de utilização em um cenário real, como o escopo da presente tese.

O contexto histórico mostra que os modelos são construídos baseados em paradigmas existentes e, como tal, são passíveis de contestação e evolução. Este cenário mostra-se ainda mais proeminente quando da modelagem de entes cujos processos são altamente atuantes, como é o caso das organizações de saúde.

Modelar os **centros cognitivos** inter-relacionados quando da assistência hospitalar, bem como a observação fenomenológica, característica de suas inter-relações, é uma atividade essencialmente paralela (Riano, Prado e Pascual, 2002). Neste contexto, a representação computacional de tais estruturas abrangem as metodologias de sistemas multiagente, que permitem um alto de grau de paralelismo, além da distribuição de tarefas entre diferentes unidades computacionais e a distribuição dos processos que referenciem situações de cognição entre diferentes entes (agentes).

No aspecto tecnológico, a aplicação do *framework* UIMA disponibiliza solução passível de expansão e utilização em diferentes realidades (hospitais).

O índice de acerto de 83,9% do mapeamento automático é compatível com o Kappa de 89%, quando da comparação do mapeamento manual entre dois diferentes especialistas, indicando que a estratégia de mapeamento automático apresenta resultado bastante similar ao mapeamento manual.

7.2 Trabalhos futuros

A grande maioria dos trabalhos que utiliza textos livres o faz com textos jornalísticos. Poucos são os que propõem técnicas para o uso de documentos ‘reais’, com elevada prevalência de inconsistências lingüísticas e formais, independente do cenário de aplicação. Neste cenário, apesar dos esforços enunciados no presente trabalho, muitas são as dimensões que podem ser exploradas, em especial na correção ortográfica e na identificação e expansão automática de acrônimos.

Para o uso em cenário real, novos agentes devem ser disponibilizados, especialmente os de observação fenomenológica (processamento de laudos de imagem, processamento de voz, etc).

Apesar de não fazer parte do escopo inicial da tese, ao longo do desenvolvimento da mesma, foi possível identificar que a área de suporte e apoio à decisão pode ser fortemente beneficiada com os resultados alcançados, neste sentido, um conjunto de agentes foi desenvolvido objetivando o inter-relacionamento do processamento de linguagem natural, mapeamento para uma estrutura ontológica e aplicação de outras estratégias de apoio e representação de conhecimento, os mesmos não objetivam esgotar o assunto, mas permitir a aplicação da presente em diferentes cenários de aplicação. O tópico em questão

provavelmente será desenvolvido em uma dissertação de mestrado, no transcurso de 2010 e 2011.

Na estratégia de mapeamento, a identificação de contextos de negação apresenta-se como elemento importante de melhoria, permitindo a aplicação de novas estratégias de mapeamento e correto mapeamento nos casos que assim se apresentem.

No contexto da desambiguação, novas estratégias podem ser propostas e validadas, especialmente com a utilização de outros tesauros. Nesta dimensão, propõe-se a especificação de um tesauro que inclua relações semânticas de desambiguação na sua estrutura. Por exemplo: dois elementos do tesauro são conectados por uma relação semântica conforme uma dada etiqueta morfológica.

A identificação de contexto é um desafio que merece melhor investigação, especialmente pelas diversas áreas envolvidas na construção do saber médico. Considerando a prevalência de sentenças curtas, formadas por poucas palavras, não oferece conteúdo suficiente para que uma análise do contexto seja efetuada para possibilitar a aplicação de técnicas convencionais de desambiguação. Utilizar um *corpus* que represente o perfil do usuário/domínio para treinar um desambiguador pode melhorar o desempenho do mapeamento e diminuir os erros oriundos de ambigüidades. Como instância do cenário pontuado, para um leigo o mapeamento e/ou busca por gordura (#fat, #lipid), deve retornar artefatos relacionados ao excesso de peso (#fat). No mesmo cenário, para um farmacêutico, o mapeamento correto estará relacionado, muito provavelmente, ao componente químico (#lipid).

Apesar dos esforços no aprimoramento do léxico do sistema MorphoSaurus, que permitiram o uso do sistema, inicialmente voltado para cenários de indexação, para o cenário da presente tese, muitas são as oportunidades de melhoria nesta dimensão. Dentre as identificadas, a que potencialmente agregará melhores resultados é a segregação dos elementos do léxico em um número maior de camadas. Atualmente, o léxico pode ser representado por dois grandes grupos de elementos, as relevantes para a indexação (*for Indexing*) e os não relevantes (*Stop-Words*), conforme apresentado na Figura 40.

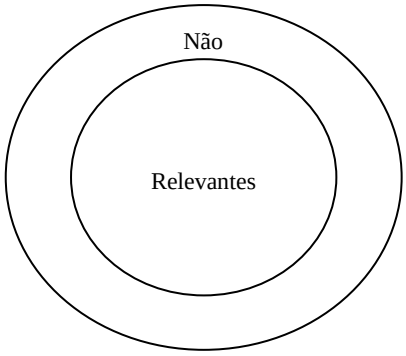


Figura 40 — Organização atual do léxico

As novas camadas classificariam os termos com relevância para indexação em três grupos distintos: os modificadores, os termos gerais do idioma e os termos específicos do domínio médico, conforme o apresentado na Figura 41.

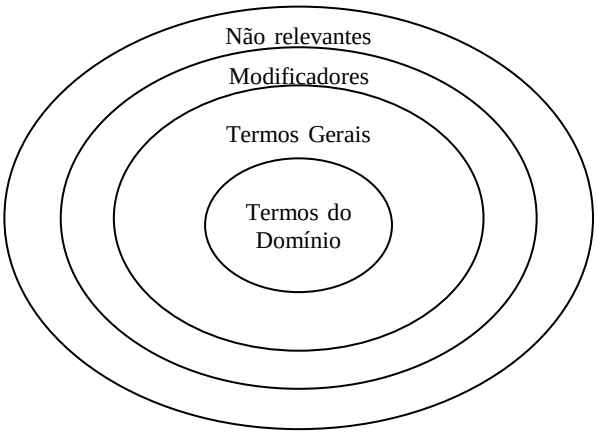


Figura 41 — Proposta de organização do léxico em 4 camadas

Na categoria de elementos não relevantes para a indexação e recuperação de documentos (*stop words*), encontram-se os pronomes pessoais, verbos auxiliares, alguns prefixos, sufixos de derivação. Este grupo já existe na atual versão do léxico. Os modificadores são os termos com significado e contexto discriminativo apenas local, ou seja, dependem de outras palavras e nunca constituem por si um elemento de relevância, em especial no domínio de Recuperação de Informação. Termos gerais são aqueles elementos de linguagem que não podem ser atribuídos a nenhuma terminologia ou domínio específico, como a maioria dos verbos e substantivos que são normalmente encontrados em um léxico. Os candidatos a termos de uso geral ou modificadores poderiam ser obtidos através da indexação dos documentos da WordNet (PRINCETON UNIVERSITY, 2009). Os modificadores podem ser identificados neste grupo pelo uso de listas de freqüência aplicadas em documentos de domínio (os próprios prontuários). Termos específicos do domínio

geralmente são substantivos que só podem ser encontrados em um léxico de determinado domínio. Para o domínio médico, os primeiros candidatos a elementos componentes do núcleo do léxico podem ser obtidos através do processamento do UMLS.

A SNOMED CT não é a única ontologia médica disponível, apesar de ser a mais completa. Neste sentido, a utilização da metodologia com outras estruturas, apresenta-se como dimensão de interesse. No caso do Brasil destaca-se a iniciativa da Terminologia Unificada da Saúde Suplementar como dimensão de interesse científico.

A aplicação de outras estratégias de PLN, bem como a observação do uso das ferramentas disponibilizadas, em um cenário dinâmico, integrado com um SI hospitalar, apresenta-se como fronteiras de pesquisas promissoras.

A adoção da SNOMED CT pelo Ministério da Saúde - MS como terminologia única para a documentação clínica mostra-se como oportunidade única para a integração do Brasil ao esforço mundial de padronização, na mesma direção, permite a revisão das diversas estratégias idiossincráticas de identificação de procedimentos e exames, permitindo que, no médio prazo, diminua-se o esforço documental voltado para fins de faturamento, permitindo a melhoria da qualidade da documentação médico-clínica.

No caso da adoção da SNOMED CT em âmbito nacional, a integração dos agentes disponibilizados no presente trabalho às diversas ferramentas disponíveis, sejam as disponibilizadas pelo MS ou as comerciais, permitirá uma rápida transição entre os modelos de codificação atuais e a nova terminologia.

As não conformidades dos mapeamentos demandam análise manual e refinamento das heurísticas aplicadas. Nesta dimensão, novas técnicas de PLN são necessárias para a melhora na qualidade do processo de mapeamento.

7.3 Conclusões

No presente trabalho, propõe-se e discutem-se metodologias para mapeamento de informações codificadas em registros clínicos para uma ontologia de domínio, a SNOMED CT, bem como os resultados auferidos pela aplicação da mesma em um domínio real (conjunto de narrativas clínicas do HCPA).

Como parte das atividades de mapeamento, os textos clínicos utilizados no experimento foram manualmente analisados, permitindo a identificação de relevante prevalência de erros nos textos, caracterizados nas mais diversas dimensões, e de importante presença de acrônimos, característica do tipo de documento.

Os textos foram pré-processados com rotinas de correção ortográfica e expansão de acrônimos, atividades necessárias para a identificação das classes gramaticais de cada *token* e a subseqüente identificação das NPs. O resultado do pré-processamento foi satisfatório, permitindo a constituição de base de treinamento, no domínio médico, para ferramentas de PLN.

Com base na técnica de *Active Learning*, foi constituído modelo, junto a OpenNLP, para etiquetagem de documentos médicos. A precisão do etiquetador auferida nos experimentos foi de 93.67%.

A identificação de frases nominais, elemento de entrada para o mapeamento, foi realizada com base na identificação de freqüências de NPs, em uma base da SNOMED CT originalmente em espanhol, na qual foram aplicadas regras específicas de compatibilização idiomática (para o português).

O mapeamento das narrativas clínicas para a SNOMED CT foi realizado com a construção de um repositório contendo a base de relações, bem como o vocabulário terminológico, processado com o sistema MorphoSaurus. O mapeamento é realizado para o conjunto de MIDs de cada NP, que passa a ser entendido como uma consulta. O conjunto de NPs e seus conceitos candidatos são ranqueados, com base na importância do MID na estrutura de termos que compõem o conceito, bem como na inter-relação entre conceitos.

O padrão ouro de validação foi constituído manualmente por especialistas de domínio o índice de acerto foi de 83,9%, quando a avaliação foi realizada por um dos especialistas que não teve acesso ao documento em qualquer uma das etapas anteriores e de 78,2%, quando a validação é realizada diretamente com base no padrão ouro (sem margem à interpretação).

Por fim, os experimentos foram organizados e codificados como agentes de software, utilizando o *framework* UIMA, permitindo, metodologicamente, que a tecnologia desenvolvida possa ser utilizada em outras soluções e, estruturalmente, possa ser aplicada em cenários de aplicação reais, caracterizados pelo paralelismo e forte demanda de recursos computacionais.

7.4 Considerações finais

Muitos são os recursos (sejam humanos ou financeiros) relacionados à saúde e, conseqüentemente, ao saber médico. Apenas no Brasil, para o exercício de 2009, estão previstos R$ 59 bilhões para custeio do Ministério da Saúde, de longe o ministério com reserva do maior montante. Apenas para estabelecimento de comparações, o segundo

ministério em recursos reservados é o da Defesa, com aproximadamente R$ 6.7 bilhões[25]. De forma complementar, é importante destacar que os procedimentos relacionados à área de saúde (consultas, exames, internações, cirurgias,...) variam de poucos reais (ex.: teste rápido para detecção de infecção pelo HIV, com o valor de R$ 1,00) a milhares de reais (ex.: transplante alogênico de células-tronco hematopoéticas de sangue perifério – não aparentado, com o valor de R$ 71.602,00).

Neste cenário, a área da assistência à saúde tem sido fortemente marcada, desde a década de 90, por uma crescente preocupação com o estímulo ao uso e efetiva utilização de boas práticas endossadas pelo conhecimento científico corrente, apoiadas em tecnologias da informação, em especial as que apóiam o processamento computacional adequado do prontuário eletrônico, com a perspectiva de melhoria da qualidade da assistência, em busca da integralidade, bem como a alocação mais eficiente de recursos, comumente limitados.

O mapeamento de narrativas clínicas para estruturas formais de representação de conhecimento, como ontologias, representará a próxima fronteira na busca pela qualidade e controle na área da saúde do século XXI.

[25] Dados do ministério do planejamento, disponíveis em: http://www.planejamento.gov.br

8 REFERÊNCIAS BIBLIOGRÁFICAS

Abney, S. **Parsing by chunks. Principle-Based Parsing**. Kluwer, p. 257-278, 1991.

ABRALAPAC. **Tabelas de Procedimentos**. Disponível em <http://www.abralapac.org.br/v2/convenios.php>. Acesso em 15 de novembro de 2009.

Adelinde, U., & Weyns, D. **Multi-Agent Systems: Simulation and Applications (Computational Analysis, Synthesis, and Design of Synamic Models)**. ISBN 978-1420070231, Los Angeles, Estados Unidos, 2009.

Aliança Saúde. **Aliança Saúde - Histórico**. Disponível em <http://www.pucpr.br/saude/alianca/>. Acesso em 15 de novembro de 2009.

Alonso-Calvo, R., Maojo, V. **An agent- and ontology-based system for integrating public gene, protein, and disease databases**. Journal of Biomedical Informatics, p. 17-29, 2009.

Aluisio, S., Pinheiro, G., Finger, M. **The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation**. CORPUS LINGUISTICS, Lancaster, Inglaterra, p. 4-21, 2003.

Ammon, D., Hoffmann, D., & Jakob, T. F. **Developing an architecture of a knowledge-based electronic patient record**. 30ª International Conference on Software Engineering, p. 653-660, 2008.

Andrade, R. **Detecção de erros em tesauro médico multilíngue através de corpora comparáveis**. 2007. 113f. Dissertação (Mestrado Ciências / Engenharia Biomédica) – Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná, Curitiba, 2007.

Andrade, R., Pacheco, E., Nohama, P., Schulz, S. **Corpus-based Error Detection in a Multilingual Medical Thesaurus**. MEDInfo 2007, Brisbane, Australia, p. 529-532, 2007.

Andrews, J. **Comparing heterogeneous SNOMED CT coding of clinical research concepts by examining normalized expressions**. Journal of Biomed Informatics, p. 1062-1069, 2008.

Andrews, J., Richesson, R., Krischer, J. **Variation of SNOMED CT coding of clinical research concepts among coding experts**. Journal of the American Medical Informatics Association, p. 497-506, 2007.

Ao, H., Takagi, T. **ALICE: an algorithm to extract abbreviations from MEDLINE**. Journal of the American Medical Informatics Association: JAMIA, v. 12, n. 5, p. 576-586, 2005.

APACHE. **What Is Lucene?** Disponível em: <http://lucene.apache.org>. Acesso em 15 novembro de 2009.

Arampatzis, A., Weide, T. **Linguistically-motivated Information Retrieval**. Encyclopedia of Library and Information Science, v.69, p.201-222, 2000.

Aronson, A. **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program**. AMIA 2001, p.17-21, 2001.

Baeza-Yates, R., Ribeiro-Neto, B. **Modern Information Retrieval**. New York: ACM Press / Addison-Wesley, 1999.

Bashyam, V., Taira, R. **Identifying Anatomical Phrases in Clinical Reports by Shallow Semantic Parsing Methods**. Computational Intelligence and Data Mining - CIDM 2007. IEEE Symposium, p. 210 – 214, 2007.

Bitencourt, J. **Monitoração de procedimentos em um tesauro multilíngüe**. 2007, 93f. Dissertação (Mestrado Ciências / Tecnologia em Saúde) – Programa de Pós-graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná, Curitiba, 2007.

Bitencourt, J., Pacheco, E., Nohama, P., Schulz, S. **Thesaurus Anomaly Detection by User Action Monitoring**. MEDInfo 2007, Brisbane, Australia, p. 655-659, 2007.

Bodenreider, O., Smith, B., Kumar, A., Burgun, A. **Investigating Subsumption in SNOMED CT: An Exploration into Large Description Logic-Based Biomedical Terminologies**. Artificial Intelligence in Medicine, v. 3, p. 183-195, 2007.

Bossen, C. **Participation, power, critique: constructing a standard for electronic patient records**. 9rd Participatory design, p. 95-104, 2006.

BRASINDICE. **Brasindice Eletrônico.** Disponível em: http://www.brasindice.com.br. Acesso em 15 de novembro de 2009.

Bustamante, F., Díaz, E. **Spelling Error Patterns in Spanish for Word Processing Applications**. LREC 2006, Genoa, Italy, p. 23-36, 2006.

Carneiro, M., Leão, B., Pereira, L. **Medical documentation: keeping and handling medical files**. Rev Soc Cardiol Estado de São Paulo, p. 832-844, 2002.

Carone, F. **Morfossintaxe**. ITAPECERICA, Atica, 2001.

Charniak, E., Hendrickson, C., Jacobson, N. **Equations for part-of-speech tagging**. 11ª National Conference on Artificial Intelligence, Menlo Park, CA, p. 784–789, 1993.

Chu, H., & Rosenthal, M. **Search engines for the World Wide Web: A comparative study and evaluation methodology**. ASIS Annual Conference Proceedings, p. 127-135, 1996.

Cimiano, P. **Ontology Learning and Population from Text: Algorithms, Evaluation and Applications**. ISBN-13: 978-0387306322, ed. Frankfurt: Springer, 2006.

CNS. **TUSS**. Disponível em <http://tuss.org.br/tuss.php>. Acesso em 15 de novembro de 2009.

Cohen. **Coefficient of agreement for nominal scales**. Educational and Psychological Measurement, p. 37-46, 1960.

Columbia University. **MedLEE - A Medical Language Extraction and Encoding System**. Disponível em <http://lucid.cpmc.columbia.edu/medlee/>. Acesso em 15 de novembro de 2009.

Costa, A., Bittencourt, G. **A Cooperation Language for Cognitive Multi-Agent Systems**. 5ø International Conference Artificial Intelligence and Symbolic Computation Theory, Implementations and Applications, p. 50-62, 2000.

Darren, G., Allen, C. **Machine translation-supported cross-language information retrieval for a consumer health resource**. AMIA 2003, p. 564–568, 2003.

Das, G., Gunopulos, D., Koudas, N. **Answering Top-k Queries Using Views**. VLDB 2006, p. 451-462, 2006.

Daumke, P., Pacheco, E., Schulz, S., Nohama, P. **Subword-based Semantic Retrieval of Clinical and Bibliographic Documents**. Methods of Information in Medicine. Ed. Schattauer. Artigo aceito para publicação.

Dawit, M., Davidsson, P. **Scalability in Distributed Multi-Agent Based Simulations: The JADE Case**. 2ª Future Generation Communication and Networking Symposia, p. 93-99, 2008.

Déjean, H., Gaussier, E. **An approach based on multilingual thesauri and model combination for bilingual lexicon extraction**. 19ª international conference on Computational linguistics. Taipei, Taiwan, p. 1-7, 2002.

Dick, R. **The computer-based patient record: an essential technology for health care**. Washington, DC, USA. National Academy Press, 1991.

Dorileo, E. A., Ponciano, M., Costa, T. **Estruturação da Evolução Clínica para o Prontuário Eletrônico do Paciente**. CBIS - X Congresso Brasileiro de Informática em Saúde, Florianópolis, 2006.

Eichelberg, M. T. **A survey and analysis of Electronic Healthcare Record standards**. ACM Computing Surveys, p. 277-315, 2005.

Ertmer, A., Ückert, F. **User Acceptance of and Satisfaction with a Personal Electronic Health Record**. London: IOS Press, 2005.

Faulconer, E., Lusignan, S. **An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar**. Informatics in Primary Care, p. 243–254, 2004.

Ferreira, A. **Dicionário Aurélio Eletrônico. Século XXI.** 3.0ª ed: Nova Fronteira, 1999.

Fiszman, M., Haug, P. **Using medical language processing to support real-time evaluation of guidelines**. AMIA 2000, p. 235-234, 2000.

Franz, A. **Independence assumptions considered harmful**. ACL 35/EACL 8, p. 182–189, 1997.

French, J., Powell, L., Gey, F., Perelman, N. **Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness**. CIKM 2001, p. 199-206, 2001.

Friedman, C., Knirsch, C., Shagina, L. **Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries**. AMIA 1999, p. 256-260, 1999.

Friedman, C., Kra, P., Rzhetsky, A. **Two biomedical sublanguages: a description based on the theories of Zellig Harris**. Journal of Biomedical Informatics, p. 222 – 235, 2002.

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

Fung, P. **A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora**. P. Fung, Parallel Text Processing, Springer, p. 1-17, 1998.

Gaines, B., Mildred, S. **Collaboration through Concept Maps**. Computer Support for Collaborative Learning, p. 135-138, 1995.

George, F. **Artificial Intelligence: Structures and Strategies for Complex Problem Solving**. ISBN-13: 978-0201648669. New York: Addison Wesley, 2001.

Gey, F. C., Kando, N. P. **Cross-language information retrieval: A research roadmap**. SIGIR Forum, p. 72–80, 2002.

Girju, R. **Out-of-context noun phrase semantic interpretation with cross-linguistic evidence**. CIKM 2006, p. 268-276, 2006.

Gomes, D. J. **Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network**. VLDB 2006, p. 415-426, 2006.

Gruber, T. **A Translation Approach to Portable Ontology Specifications**. Knowledge Acquisition (Special issue: Current issues in knowledge modeling), v. 5, n. 2, p. 199-200, 1993.

Grundstein, M. **Knowledge Engineering Within the Company: An Approach to Constructing and Capitalizing the Knowledge Assets of the Company**. IAKE 1992, p. 16-19, 1992.

Guarino, N. **Formal Ontology in Information Systems**. FOIS 1998, Trento, Italy, p. 3-15, 1998.

Gundersen, M., Haug, P., Pryor, T. **Development and evaluation of a computerized admission diagnoses encoding system**. Computer Biomed Res, p. 351-372, 1996.

Hahn, U., Wermter, J. **Tagging Medical Documents with High Accuracy**. AI Specific Application Areas - Natural Language Processing. Berlin: Springer Berlin / Heidelberg - Lecture Notes in Computer Science, p. 852-861, 2004.

Hahn, U., Daumke, P., MARKÓ, Schulz, S., Nohama, P. **Multilingual MESH Mapping**. MEDINFO 2004, San Francisco, p. 60-66, 2004.

Halliday, M. **Current Ideas in Systemic Practice and Theory**. London: Pinter, 2001.

Harris, Z. **A Theory of Language and Information: A Mathematical Approach**. Oxford University Press, USA, 1991.

Hayes-Toth, B. **An Architecture for Adaptive Intelligent Systems**. Arificial Intelligence: Special Issue on Agents and Interactivit , p. 72-76, 1995.

HCPA. **Apresentação Institucional**. Disponível em <http://www.hcpa.ufrgs.br/downloads/institucional/Apresentacao_institucional.pdf>. Acesso em 15 de novembro de 2009,

Hedlund, T., Kalervo, J. **Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language retrieval**. Information Processing & Management , p. 147–161, 2001.

Hersh, W., Buckley, C. **Ohsumed: An Interactive Retrieval Evaluation and new large Test Collection for Research**. 17ª ACM SIGIR Conference on Research and Development in Information Retrieval, p. 192-201, 1994.

Hersh, W., Mailhot, M., Arnott-Smith, C. **Selective automated indexing of findings and diagnoses in radiology reports**. Journal of Biomedical Informatics, p. 262–273, 2001.

Honeck, M., U., H., Klar R., S. **Text Retrieval Based on Medical Subwords**. MIE2002. IOS Press, p. 241-245, 2002.

Huang, Y., & Lowe, H. J. **Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon**. Journal of the American Medical Informatics Association, p. 275 – 285, 2005.

Huhns, M., Munindar, P. **Readings in Agents**. New York: Morgan Kaufmann, Inc, 1998.

Huhns, M., & Singh, M. **Agents**. New York: Morgan Kaufmann, 1997.

IHTSDO. **International Health Terminology Standards Development Organisation**. Disponível em: <http://www.ihtsdo.org/snomed-ct>. Acesso em 15 de novembro de 2009.

Iivonen, M. **Searchers and Searchers: Differences between the Most and Least Consistent Searchers**. 18ª ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Seattle, Washington, USA, p. 149-157, 1995.

Jackson, P., Moulinier, I. **Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization**. Amsterdam. John Benjamins Publishing Co, 2002.

Jacquemin, C., Klavans, J. **Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax**. 35º Annual Meeting of the Association for Computational Linguistic (ACL), 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, p. 25-74, 1999.

Jeschke, S., Natho, N., Wilke, M. **mArachna - Applying Natural Language Processing Techniques to Ontology Engineering**. 7ª IEEE International Conference on Advanced Learning Technologies, p. 571-575, 2007.

Jiang, G., Chute, C. **Auditing the semantic completeness of SNOMED CT using formal concept analysis**. Journal of the American Medical Informatics Association, n. 1, p. 89-102, 2009.

Jones, D., Somers, H. **New Methods in Language Processing**. London, UK: University College Press, 1998.

Jurafsky, D., Martin, J. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. Chicago: Prentice Hall, 2000.

Kagolovsky, Y., Moehr, J. R. **A New Look at Information Retrieval Evaluation: Proposal for Solutions**. Journal of Medical Systems, Volume 28, n. 1, p. 17-24, 2004

Kanou, H., Jouber, M., & Maury, G. **Towards a Knowledge-based Multimedia Electronic Patient Record**. Electronic Patient Records in Medical Practice, p. 288-291, 1998.

Klück, M., GUIMARÃES, J. **Questões éticas e legais do prontuário do paciente: Da teoria a prática**. VIII CBIS - Congresso da Sociedade Brasileira de Informática em Saúde, Natal, RN, p. 15-22, 2002.

Klück, M., Guimarães, J. R. **Sumário Eletrônico de Alta: garantindo a continuidade da assistência ao paciente através da informação**. Informática Pública, Belo Horizonte, p. 123-139, 1999.

Kodratoff, Y. **Machine Learning: An Artificial Intelligence Approach**. San Francisco, USA: Morgan Kaufmann Publishers Inc, v. 1, 1990.

Konder, L. **A revanche da dialética**. São Paulo: UNESP, 2002.

Kusniercyk, W. **Nontological Engineering. Proceedings of the International Conference on Formal Ontology in Information Systems**, v. 11, p. 39-50, 2006.

LácioWeb. **LácioWeb**. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/downloads.htm>. Acesso em 15 de novembro de 2009.

Letrilliart, L., Viboud, C., Boëlle, P. F. **Automatic coding of reasons for hospital referral from general medicine free-text reports**. AMIA 2000, p. 487–491, 2000.

Lewis, A. **Health informatics: information and communication**. Advances in Psychiatric Treatment, p. 165-171, 2002.

LOINC. **Logical Observation Identifiers Names and Codes**. Disponível em <http://loinc.org/>. Acesso em 15 de novembro de 2009.

Long, W. **Lessons extracting diseases from discharge summaries**. AMIA 2007, p. 478-482, 2007.

Lovis, C., Baud, R., Planche, P. **Power of expression in the electronic patient record: structured data or narrative text**? International Journal of Medical Informatics, v. 58, p. 101-110, 2000.

Lussier, Y., Shagina, L., Friedman, C. **Automating SNOMED coding using medical language understanding: a feasibility study**. AMIA 2001, p. 418-422, 2001.

Maes, P. **Artificial Life Meets Entertainment: Life like Autonomous Agents**. Communications of the ACM, v.38, p. 11-21, 1995.

Manning, C., Schuetze, H. **Foundations of Statistical Natural Language Processing**. Cambridge, USA: The MIT Press, 1999.

Marchiori, M. **The quest for correct information on the Web: Hyper search engines**. Computer Networks and ISDN Systems, v. 29, p. 1225-1236, 1997.

Marcus, M. **A Theory of Syntactic Recognition for Natural Language**. Cambridge, USA: MIT Press, 1980.

Markó, K., Daumke, P., Schulz, S. **Large-scale evaluation of a medical cross-language information retrieval system**. Studies in health technology and informatics, v. 129, p. 392-396, 2007.

Markó, K., Schulz, S., Hahn, U. **Cross-Language MeSH indexing using morphsemantic normalization**. AMIA 2003, p. 40-47, 2003.

Marques, A. **Para entender as linguagens documentárias**. 2ª Edição. São Paulo: Polis, 2002.

Maryam, A., Leidner, D. **Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues**. MIS Quarterly, v. 25, p. 107-136, 2001.

Massad, E., Marin, H. **O prontuário eletrônico do paciente na assistência, informação e conhecimento médico**. Revista TEXTOS de la CiberSociedad, v. 11, p. 1-20, 2003.

McBryan, O. **GENVL and WWWW: Tools for Taming the Web**. 1ª International World Wide Web Conference, CERN, Geneva, p.15-19, 1994.

Mealey, G. **Another Look at Data**. FJCC, p. 525-534, 1967.

Michalsky, R., & Kodratoff, Y. **Machine Learning: An Artificial Intelligence Approach**. New York: Morgan Kaufmann, v. 3, 1990.

Miettinen, M., & Korhonen, M. **Information Quality in Healthcare: Coherence of Data Compared between Organization's Electronic Patient Records**. 21ª IEEE International Symposium on Computer-Based Medical Systems, p. 488-493, 2008.

Minchin, R., & Vangenot, C. **Ontology for Mapping Diagnostic Knowledge Systems**. 19ª IEEE Symposium on Computer-Based Medical Systems, p. 593-598, 2006.

Minsky, M. **A framework to represent knowledge**. The Psychology of Computer Vision, McGraw-Hill, p. 211-277, 1975.

Minsky, M. **The Society of Mind Simon and Schuster**. New York, 1986.

MITA. **Medical Imaging and Technology Alliance.** Disponível em: <http://medical.nema.org>. Acesso em 15 de novembro de 2009.

Moghrabi, C., Moussa, S., Eid, M. **Natural Language Processing Complexity and Parallelism**. 16ª Annual International Symposium on High Performance Computing Systems and Applications, p. 276-277, 2002.

Moorman, P., Musen, M. **Electronic patient records in medical practice: a multidisciplinary endeavor**. Methods of Information in Medicine, v. 38, 287-288, 1999.

Morin, E. **Introdução ao pensamento complexo**. 3ª edição. Lisboa: Instituto PIAGET. 2001.

Morton, T. **Using semantic relations to improve information retrieval**. University of Pensilvania: MIT Press, 2006.

Moshchuk, A., & Bragin, T. **A Crawler-based Study of Spyware in the Web**. Annual Network and Distributed System Security Symposium. San Diego, 2006.

Murphy, K. **In praise of Bayes**. Computational Intelligence, v. 4, p. 92-93, 1988.

Nascimento, M. E. **String Processing and Information Retrieval**. 10ª International Symposium, SPIRE 2003, Manaus, Brazil, 2003.

NLM. **United States National Library of Medicine**. Disponível em: <http://www.nlm.nih.gov/research/umls/rxnorm>. Acesso em 15 de novembro de 2009.

Novak, J. **A Theory of Education**. Ithaca. Illinois: Cornell University Press, 1977.

Noy, N., McGuinness, D. **Ontology Development 101: A Guide to Creating Your First Ontology**. 2001. Disponível em: <http://protege.stanford.edu/publications/ontology_development/ontology101.pdf>. Acesso em 15 de novembro de 2009.

Nytro, O., Sorby, I. D. **Query-based requirements engineering for health care information systems: Examples and prospects**. International Conference on Software Engineering, p. 62-72, 2009.

Okazaki, N., Ananiadou, S. **Clustering acronyms in biomedical text for disambiguation**. 5ª International Conference on Language Resources and Evaluation, p. 959–962, 2006.

OPENCLINICAL. **Knowledge Management for Medical Care**. Disponível em: <http://www.openclinical.org>. Acesso em 15 de novembro de 2009.

OPENNLP. **OpenNLP**. Disponível em <http://opennlp.sourceforge.net/>. Acesso em 15 de novembro de 2009.

Pacheco, E., Cancian, P., Nohama, P., Schulz, S. **MorphoSaurus: A Cross-Language Information Retrieval System**. ELAN, p. 30-37, 2007.

Pacheco, E., Nohama, P., Andrade, R., Schulz, S. **The MORPHOSAURUS Medical Subword Lexicon** . 5ª International Conference on Language Resources and Evaluation. LREC 2006, Genova, p. 50-62, 2006.

Pacheco, E., Stenzhorn, H., Schulz, S., Nohama, P. **Detecting Underspecification in SNOMED CT Concept Definitions Through Natural Language Processing**. AMIA 2009, Artigo aceito para publicação.

Pacheco, E., Stenzhorn, H., Schulz, S., Nohama, P. **Semantic Mapping of Clinical Narratives**. Medical Informatics Europe (MIE), p. 23-30, 2009.

Page, L., Brin, S., Motwani, R., Winograd, T. **The PageRank Citation Ranking: Bringing Order to the Web**. Stanford Digital Library Technologies Project, 1998.

Pakhomov, S. **Semi-supervised Maximum Entropy based approach to acronym and abbreviation normalization in medical texts**. ACL 2002: 40ª Annual Meeting on Association for Computational Linguistics, p. 160-167, 2001.

Paladino, V. **Analise Sintática - Teoria e Prática**. São Paulo: Freitas Bastos, 2006.

Pan, H., Han, Q., Yin, G. **A ROI-Based Mining Method with Medical Domain Knowledge Guidance**. Internet Computing in Science and Engineering, p. 91-97, 2008.

Park, J. C. **Using Combinatory Categorical Grammar to Extract Biomedical Information**. IEEE Intelligent Systems, v. 16, p. 62-67, 2001,

Parry, D. **A fuzzy ontology for medical document retrieval**. 2º workshop onAustralasian information security, Data Mining and Web Intelligence, and Software Internationalization. v. 32, p. 30-38, 2004.

Patrick, T. **SNOMED CT coding variation and grouping for other findings in a longitudinal study on urea cycle disorders**. AMIA 2008, p. 11-15, 2008.

Percy, C. E., Lancashire, I. **Synchronic Corpus Linguistics**. 16ª International Conference on English Language and Research on Computerized Corpora (ICAME 16), Amsterdã, Rodipi, p. 24-36, 1996.

Pestotnik, S. **Medical informatics: Meeting the information challenges of a changing health care system**. Journal of Informed Pharmacotherapy, v. 2, p. 1-3, 2000.

Pollock, J., Zamora, A. **Automatic spelling correction in scientific and scholarly text**. Communications of the ACM, v. 27, n. 4, 358-368, 1984.

Porcheret, M., Hughes, R., Evans, D. K. **Data quality of general practice electronic health records: The impact of a program of assessments, feedback, and training**. Journal of the American Medical Informatics Association, v. 11, p. 78-86, 2004.

Porter, M. **Snowball: A Language for Stemming Algortihms**. Disponível em: <http://www.snowball.tartarus.org/texts/introduction.html>. Acesso em 15 de novembro de 2009.

Price, S., Delcambre, L., & Nielsen, M. **Using semantic components to express clinical questions against document collections**. International workshop on Healthcare information and knowledge management, p. 9-16, 2008.

PRINCETON UNIVERSITY. **WordNet**. Disponível em: <http://wordnet.princeton.edu/>. Acesso em 15 de novembro de 2009.

Pustejovsky, J., Cochran, B. K., Morrell, M. **Automatic extraction of acronym-meaning pairs from Medline**. 10º World Congress on Medical Informatics, p. 371–375, 2001.

Puustjärvi, J. **Using Knowledge Management Technologies in Searching Medicinal Learning Objects**. International Conference on eHealth, Telemedicine, and Social Medicine, p. 190-193, 2009.

Qamar, R., Rector, A. **Semantic Mapping of Clinical Model Data to Biomedical Terminologies to Facilitate Data Interoperability**. Healthcare Computing, Harrogate, U.K., p. 27-31, 2007.

Quillian, M. **Semantic memory**. Semantic Information Processing, M.I.T. Press, Cambridge, p. 215-272, 1968.

Ratnaparkhi, A. **Maximum Entropy Models for Natural Language Ambiguity Resolution**. University of Pennsylvania, Philadelphia, PA, 1998.

Ren, X., Perrault, F. **The typology of unknown words: an experimental study of two corpora**. International Conference On Computational Linguistics, v. 1, p. 408 – 414, 1992.

Riano, D., Prado, S., Pascual, A. **A multi-agent system model to support palliative care units**. Computer-Based Medical Systems, p. 35- 40, 2002.

Ringlstetter, C., Schulz, K. U., Mihov, S. **Adaptive text correction with Web-crawled domain-dependent dictionaries**. ACM - Transactions on Speech and Language Processing, p. 30-39, 2007.

Robert, C. **The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation**. New York: Springer Verlag, 2007.

Rokach, L., Maimon, O. **Data Mining with Decision Trees: Theroy and Applications (Machine Perception and Artificial Intelligence)**. London: World Scientific Publishing Company, 2008.

Russel, S., Norvig, P. **Artificial Intelligence - A Modern Approach**. Prentice Hall, Inc, 1995.

Sager, N. **Syntactic Analysis of Natural Language**. Advances in Computers 8, Academic Press, NY, p. 153-188, 1967.

Sager, N., Lyman, M. **Natural language processing and the representation of clinical data**. Journal of the American Medical Informatics Association, v. 1, p. 142-160, 1994.

Salem, A., Alfonse, M. **Ontology versus semantic networks for medical knowledge representation**. Recent Advances In Computer Engineering, p. 769-774, 2008.

Salton, G., Buckley, C. **Term Weighting Approaches in Automatic Text Retrieval**. Cornell University Ithaca, NY, USA, 1987.

Sanderson, H., Adams, T., Budden, M., Hoare, C. **Lessons from the central Hampshire electronic health record pilot project: evaluation of the electronic health record for supporting patient care and secondary analysis**. Journal of Medical Internet Research, v. 328, p. 875–878, 2004.

Sankoff, D., Kruskal, J. **Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison**. New York: Addison-Wesley, 1983.

Sardinha, T. **Lingüística de Corpus**. São Paulo: Manole, 2004.

Sayão, L. **Bases de dados: a metáfora da memória científica**. Ciência da Informação, p. 1-6, 1996.

Schulz, K., Mihov, S. **Fast String Correction with Levenshtein-Automata**. International Journal of Document Analysis and Recognition, p. 67-85, 2002.

Schulz, S., & Hahn, H. **Syntatic and Semantic Aspects of Subword Indexing**. IJMI, Italy, p. 42-47, 2006.

Schulz, S., Hahn, U. **Biomedical Text Retrieval in Languages with a Complex Morphology**, ACL 2002 - Workshop on Natural Language Processing in the Biomedical Domain, p. 61-68, 2002.

Schulz, S., Han, U. **Morpheme-based cross-language indexing for medical document retrieval**. International Journal of Medical Informatics, p. 87-99, 2000.

Schulz, S., Johansson, I. **Continua in Biological Systems**. The Monist, p. 499 – 522, 2007.

Schulz, S., Romacker, M., Hahn, U. **Part-whole reasoning in medical ontologies revisited – introducing SEP triplets into classification-based description logics**. AMIA 1998, p. 830-834, 1998.

Schulz, S., Sbrissia, E., Nohama, P. **Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon**. 20ª International Conference on Computational Linguistics, Geneva – Swiss, 2004.

Schulz, S., Stenzhorn, H., Boeker, M., Smith, B. Vantagens e Limitações das Ontologias Formais. Revista Eletrônica de Comunicação, Informação e Inovação em Saúde (RECIIS), v. 12, p. 15 – 22, 2009.

Schwartz, A. S., Hearst, M. A. **A simple algorithm for identifying abbreviation definitions in biomedical text**. Proceedings of the Pacific Symposium on Biocomputing 2003, p. 451–462, 2003.

Severino, A. **Síntese do Conhecimento**. AEC 31, p. 9-30, 2002.

Shmeil, M. A. **Sistemas Multiagente na Modelagem da Estrutura e Relações de Contratação de Organizações**. Porto: Faculdade de Engenharia da Universidade do Porto, 1999.

Silva, F., Tavares-Neto, J.**Avaliação de Prontuários Médicos de Hospitais de ensino do Brasil**. Revista Brasileira de Educação Médica. v.31, n. 2, p. 113-126, 2007.

Simon, H. **The Sciences of The Artificial. Massachusetts**, The M.I.T. Press, p. 1-22, 1968.

Sinclair, J. **From Theory to Practice. Spoken english on computer: transcription, mark-up and applicaton**. Londres: Logman, 1995.

Sivagurunathan, K., Chountas, P. **Representation & modeling of electronic patient records**. International conference on Computational methods in sciences and engineering, p. 598-603, 2003.

Skov, B. **Supporting information access in a hospital ward by a context-aware mobile electronic patient record**. Personal and Ubiquitous Computing, v.5, n. 2, p. 205-214, 2006.

Smeaton, A. **Information Retrieval: Still Butting Heads with Natural Language Processing?** SCIE-97: International summer school on information extraction, Frascati , Italy, v. 1299, p. 115-138, 1997.

Smith, D., Cypher, A., Spohrer, J. **Programming Agents without a Programming Language**. Communications of the ACM, v.37, n. 7, p. 37-44, 1994.

Snowball. **Snowball**. Disponível em <http://snowball.tartarus.org/>. Acesso em 15 de novembro de 2009.

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

Sowa, J. **Knowledge Representation: Logical, Philosophical, and Computational Foundations**. Course Technology, California, Brooks/Cole, 1999.

Spyns, P. **Natural language processing in medicine: an overview**. Methods Of Information In Medicine, v.35, p. 285-301, 1996.

Stenzhorn, H., Schulz, S., Smith, B. **Adapting Clinical Ontologies in Real-World Environments**. Journal of Universal Computer Science, v.14, n.22, p. 3767-3780, 2008.

Stubbs, M. **Text and corpus analysis: computer-assisted studies of language and culture**. Oxford: Blackwell, 1996.

SUS. **DATASUS**. Disponível em <http://w3.datasus.gov.br/datasus/datasus.php>. Acesso em 15 de novembro de 2009.

Tempich, C., Staab, S. **Ontology Engineering Revisited: An Iterative Case Study**. Springer, ESWC, v. 4011, p. 110-124, 2006.

Thro, E. **The Artificial Intelligence Dictionary**. San Marcos, CA.: Microtrend Books, 1991.

Tomanek, K., Wermter, J., Hahn, U. **An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data**. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), p. 486-495, 2007.

Torii, M., Liu, H., Hu, Z. **A comparison study of biomedical short form definition detection algorithms**. ACM 1st International Workshop on Text Mining in Bioinformatics, p. 52-59, 2006.

Tretiakov, A., Hunter, I., Whidett, D., Sutinen, E. **Coding of Medical Records via Restrictive Semantic Topic Tracking**. Health Care and Informatics Review, p. 101-108, 2006.

UMC. **Uppsala Centre for International Drug Monitoring**. Disponível em: <http://www.umc-products.com>. Acesso em 15 de novembro de 2009.

Veras, C. M., Martins, M. S. **A confiabilidade dos dados nos formulários de Autorização de Internação Hospitalar (AIH)**. Revista de Saúde Pública, v. 32, n.6, 1994.

Veroyatnostei, S. T., Statistika, M. **Automatic error correction in inflected languages**. Journal of Mathematical Sciences, p. 2263-2279, 2005.

Villavicencio, A., Viccari, R. **Evaluating Part-of-Speech Taggers for the Portuguese Language**. Atas do II Encontro para o processamento computacional de português escrito e falado. PROPOR 1996, p. 159-168, 1996.

Vosse, T. **Detecting and correcting morpho-syntactic errors in real texts**. 3ª conference on Applied natural language processing, p. 111 – 118, 1992.

Voutilainen, A. **NPtool, a detector of English noun phrases**. Workshop on Very Large Corpora, p. 48-57, 1995.

Walczak, S. **A Multiagent Architecture for Developing Medical Information Retrieval Agents**. Journal of Medical Systems, v. 27, n. 5, p.52-64, 2003.

Waltz, D. **An Opinionated History of AAAI**. AI Magazine, p. 45-47, 2005.

Wang, Y., Halper, M., Min, H., Perl, Y., Chen, Y., Spackman, K. **Structural methodologies for auditing SNOMED**. Journal of Biomedical Informatics, v. 5, n. 40, p. 561-581, 2007.

Waterman, D. **A Guide to Expert System**. Addison-Wesley Publishing Company, USA, 1986.

Weaver, R. **Resistance to computer innovation: knowledge coupling in clinical practice**. SIGCAS 2002, p. 16-21, 2002.

Weber-Jahnke, J., Price, M. **Engineering Medical Information Systems: Architecture, Data and Usability & Security**. 29th International Conference on Software Engineering, p. 188-189, 2007.

Weischedel, R., Meteer, M., Schwartz, R. **Coping with ambiguity and unknown words through probabilistic models**. Computational Linguistics, p. 359–382, 1993.

Weiss, G. **Multiagent systems: a modern approach to distributed artificial intelligence**. Massachustes: MIT Press, 1999.

WHO. **World Health Organization**. Disponível em: <http://www.who.int/classifications/icd>. Acesso em 15 de novembro de 2009.

WHOCC. **WHO Collaborating Centre for Drug Statistics Methodology**. Disponível em: <http://www.whocc.no/atcddd>. Acesso em 15 de novembro de 2009.

Wille, S., Bruza, P. **Users Model of the Information Space: the Case for Two Search Models**. SIGIR FORUM 1995, p. 205 – 211, 1995.

Win, K., Susilo, W. **Personal Health Record Systems and Their Security Protection**. Journal of Medical Systems, p. 309-315, 2006.

Wong, W., Liu, W., Bennamoun, M. **Enhanced Integrated Scoring for Cleaning Dirty Texts**. IJCAI-2007 - Workshop on Analytics for Noisy Unstructured Text Data, Hyderabad, India, p. 55-62, 2007.

Woods, W. **What's in a link: Foundations for semantic networks**. Representation and Understanding: Studies in Cognitive Science. New York: Academic Press, 1975.

Wooldridge, M. **Intelligent Agents: Theory and Practice**. Knowledge Engineering, p. 115-152, 1995.

Wooldridge, M., Jennings, N. **Intelligent Agents: Theory and Practice**. The Knowledge Engineering Review, 1994.

Wren, J. D., Chang, J. T. **Biomedical term mapping databases**. Nucleic Acids Research, p. 289–293, 2005.

Yujian, L., Bo, L. **A Normalized Levenshtein Distance Metric**. Pattern Analysis and Machine Intelligence, IEEE Transactions, v. 29, n. 6, p. 1091-1095, 2007.

Zweigenbaum, P. **Natural Language Processing in the Medical and Biological Domains: a Parallel Perspective**. 3rd International Symposium on Semantic Mining in Biomedicine, Turku, Finland, p. 3-4, 2008.

Anexos

9 ANEXO A – CÁLCULO DE SIMILARIDADE

No modelo de espaço vetorial os documentos, bem como as consultas aplicadas sob estes documentos, são representadas como vetores de termos, sendo que por termos entendem-se como ocorrências únicas no conjunto de documentos.

Os documentos recuperados como resultados do processamento de uma consulta são representados similarmente (através do cálculo de similaridade). Para cada termo, em uma consulta consultas são atribuídos pesos que determinam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por estes vetores dá-se o nome de ɵ. Pelo *cos*(ɵ) determina-se a proximidade da ocorrência. O cálculo da similaridade é determinado pelo ângulo entre os vetores que representam o par documento e consulta, através da seguinte fórmula(Salton e Buckley, 1987).

$$sim(\vec{d},\vec{q}) = \frac{\vec{d} \bullet \vec{q}}{\left|\vec{d}\right| * \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{d,i} w_{q,i}}{\sqrt{\sum_{i=1}^{t} w^2_{d,i}} \sqrt{\sum_{i=1}^{t} w^2_{q,i}}}$$

Equação 14

Onde, |d| é o módulo do vetor d * cos(ɵ) é o cosseno do ângulo formado pelos vetores que representam os dois documentos d e q, esse valor varia entre 0 e 1, sendo que 1 indica maior similaridade e 0 total dissimilaridade.