Universidade Federal de Santa Catarina
Programa de Pós-graduação em
Engenharia de Produção

Fábio Alexandrini

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

Tese de Doutorado

FLORIANÓPOLIS
2005

Fábio Alexandrini

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

Tese apresentada ao Programa de Pós-graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, como requisito parcial para obtenção do título de Doutor em Engenharia de Produção.

Orientador: Prof. Dr. rer.nat. Aldo von Wangenheim

FLORIANÓPOLIS

2005

Ficha catalográfica

A382d Alexandrini, Fábio.

Desenvolvimento de uma metodologia de interpretação, recuperação e codificação inteligente de laudos médicos independente de idioma / Fábio Alexandrini, Aldo von Wangenheim.– Florianópolis, 2005.

100f. : figs. ; 30 cm.

Tese (doutorado) – Universidade Federal de Santa Catarina, Programa de Pós-graduação em Engenharia de Produção

Anexos : f. 93– 99.

Bibliografia : f. 91 – 92.

1. Inteligêcncia artificial. 2. Processamento de linguagem natural (Computação). 3. Resultados médicos. I. Von Wangenheim, Aldo. II. Título.

Elaborada por: Simone da Silva Conceição

CRB 14/526

Fábio Alexandrini

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

Esta tese foi julgada e aprovada para a obtenção do título de **Doutor em Engenharia de Produção** no **Programa de Pós-graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina.

Florianópolis, 01 de abril de 2005.

_____

Prof. Dr.Edson Pacheco Paladini

Coordenador do Programa

BANCA EXAMINADORA

_____

Prof. Dr. rer.nat. Aldo von Wangenheim
Universidade Federal de Santa Catarina
Orientador

_____

Prof.Dr. Roger Walz
Universidade Federal de Santa Catarina

_____

Prof. Dr. rer.nat. Michael M. Richter, Dr.
Universität Kaiserskautern

_____

Prof. Dr. Rogério Cid Bastos
Universidade Federal de Santa Catarina

_____

Prof. Dr. rer.nat. Eros Comunello
Universidade do Oeste de Santa Catarina

Agradecimentos

À Universidade para o Desenvolvimento do Alto Vale do Itajaí, Universidade Federal de Santa Catarina, Universität Kaiserslautern, Hospital Regional Alto Vale, DAAD e CAPES pela oportunidade da realização do meu Doutorado.

Aos Prof. Dr. rer.nat. Aldo von Wangenheim. e Prof. Dr. rer.nat. Michael M. Richter, pela orientação e por terem acreditado, estimulado e direcionado melhor meu tema.

Ao Prof. Dr. rer.nat. Eros Comunello, pelo companheirismo e o incentivo indispensáveis para a realização deste trabalho.

A todos os colegas da Universidade para o Desenvolvimento do Alto Vale do Itajaí, Universidade Federal de Santa Catarina, Universität Kaiserslautern e Hospital Regional Alto Vale, que direta ou indiretamente contribuíram com esta tese e participaram de minha vida acadêmica e profissional

A minha esposa e companheira Carla, por ter sido minha incentivadora e companheira, por existir em minha vida e estar sempre comigo.

Aos meus pais Fernandes e Lúcia, meus filhos Iago e Enzo, meus sogros Valtrudes e João e amigos pela compreensão e apoio, minha gratidão.

Ao nosso senhor Deus, Jesus Cristo e a Nossa Senhora, mestres maiores, por terem-me colocado frente a esse desafio e por terem-me dado as melhores de todas as armas para enfrentá-lo, a fé, a perseverança, a esperança e o amor.

Resumo

ALEXANDRINI, Fábio. Desenvolvimento de uma metodologia de interpretação, recuperação & codificação inteligente de laudos médicos independente de idioma 2005 (100f) Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC. Florianópolis.

A rotina médica gera diversos registros para documentar o estado de saúde dos pacientes, em papel ou em sistemas informatizados de hospitais e clínicas que, normalmente, têm maior enfoque no controle administrativo e financeiro, relegando ao segundo plano os dados do prontuário que ficam registrados em arquivos ou em campos de texto, sem estruturação. Atualmente, existem padrões de armazenamento de informações médicas que permitem estruturação e manipulação adequada dos prontuários eletrônicos de paciente que permitem a interoperabilidade de informações. Visando resgatar os registros antigos e estruturá-los em padrões de armazenamento internacionais. Este trabalho centra-se na elaboração de uma metodologia para interpretação, recuperação e codificação inteligente de laudos médicos utilizando técnicas de PLN - Processamento de Linguagem Natural combinadas com terminologias médicas internacionais. Descreve uma ferramenta de software que recupera e interpreta laudos médicos em padrão texto, baseando-se na nomenclatura SNOMED (Systematized Nomeclature of Medicine) para estruturação desses laudos visando à integração com softwares de edição de Laudos Estruturados baseado no DICOM SR-Structured Report.

Palavras Chaves: Laudos Médicos, Processamento Linguagem Natural, DICOM SR.

Abstract

ALEXANDRINI, Fábio. Desenvolvimento de uma metodologia de interpretação, recuperação & codificação inteligente de laudos médicos independente de idioma 2005 (100f) Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC. Florianópolis.

Most health care institutions have a precious legacy of clinical reports written in natural language, or simplified grammatical structure. Unfortunately content based retrieval of information from these reports is inefficient due to the peculiarities of natural language. This prevents institutions from sharing clinical records without waste of precious time and resources. The present research effort focuses on the development of methods based on knowledge about normal findings in medical reports and the Systematized Nomenclature of Medicine - SNOMED with the objective of translating the medical reports into a representation more suitable for content recovery and that can be, in further work, rendered into reports compliant to internationally accepted standards such as DICOM Structured Report. A set of medical reports, provided by a Brazilian and a German health care institution, was used as source of sample subjects for the experiment.

Key Words: Medical Reports, Natural Language Processing, DICOM SR.

Sumário

Lista de Figuras

Lista de Quadros

1 Considerações Iniciais

1.1 Introdução

Muitas instituições de saúde buscam implementar sistemas de registro clínico eletrônico visando à maior agilidade no acesso aos dados de paciente, melhorando assim o atendimento. Mas normalmente, os sistemas informatizados possuem forte enfoque administrativo e financeiro em detrimento da preocupação com as informações dos pacientes como laudos e prescrições, que são registrados apenas em arquivos de texto. A escrita em textos livres é o método mais comum, porém podem-se utilizar inúmeros recursos de linguagem, palavras sinônimas, abreviaturas, linguagem figurada, entre outros, que dificultam a interpretação dos mesmos e conduzem a erros e equívocos, principalmente quando se utiliza esta técnica em computadores, reduzindo o texto livre a apenas um conjunto de dados de difícil manipulação e interpretação.

A proposta de combinação de diversas técnicas integradas a sistemas de informações na área da saúde seria uma solução viável que permitiria o uso efetivo de sistemas de registros médicos com técnicas de recuperação de informações que, habitualmente, ficam apenas armazenadas em bases de dados ou em backup (cópia de segurança) sem utilidade para o dia-a-dia. Porém, as informações dos casos registrados até o momento e que estão em formato texto, são de extrema importância no tratamento dos pacientes, assim como para comparações com casos de novos pacientes.

Mas, para que estas possam ser recuperadas e utilizadas de forma plena pelos profissionais da área da Saúde, é necessário que estejam ordenadas e padronizadas para facilitar o entendimento, principalmente, para possibilitar agilidade na sua recuperação. Ao observar-se esta situação, podem-se vislumbrar alguns cenários possíveis de aplicação.

1.2 Cenários de Aplicação

1.2.1 Cenário #1 – Hospital Informatizado há vários anos

Considere a situação de um hospital que utiliza um sistema legado com forte enfoque administrativo e financeiro, que usa armazenamento de documentos médicos textuais há mais de 10 anos. Se este hospital possuir uma média de atendimentos de 7500 pacientes por mês (como, por exemplo, média do Hospital Regional Alto Vale pelo SUS no Ano de 2002), somente nesse período acumularia aproximadamente 900.000 registros médicos que ficariam armazenados para servir as estatísticas sobre atendimentos e ou indexado apenas por algum outro campo de interesse puramente administrativo. Ao considerar-se a recomendação do CFM de armazenar por 20 anos todos os documentos sobre o atendimento de pacientes, obter-se-ia o dobro de casos disponíveis.

Se o conteúdo destes documentos estiver apenas em formato texto, serão de pouca utilidade para os médicos na busca de casos semelhantes aos de um paciente atual, em função da dificuldade e demora nos processos de pesquisa quase que seqüencial em uma extensa base de textos, mesmo trabalhando na rede interna do próprio hospital. Considerando-se o surgimento de novos casos de difícil diagnóstico, poder-se-ia fazer uso de uma ferramenta para preparação e estruturação de laudos em formato de fácil intercâmbio que serviria como base de consulta a diagnósticos anteriores que são relevantes para o entendimento ou embasamento de decisões no novo caso em questão.

A ausência de ferramentas dessa natureza inibe a utilização e o reaproveitamento do conhecimento armazenado nas bases textuais do sistema legado, que foram geradas no decorrer dos anos, muitas vezes, por médicos especialistas que nem fazem mais parte do corpo clínico do hospital, impossibilitando, assim, qualquer tipo de discussão do caso de forma presencial. Isto, em muitas situações, pode significar maior agilidade, principalmente no tratamento e recuperação dos pacientes, por encontrar a melhor indicação do melhor caminho deixado pelo outro especialista, ou ainda evitar o emprego de tratamentos que não foram eficientes, diminuindo a necessidade de experimentações que podem

significar preciosos momentos ou até mesmo a vida do paciente, além da conseqüente redução dos custos para o hospital.

1.2.2 Cenário #2 – Consultório Médico Compartilhando Informações

Considerando que um médico possuísse um vinculo junto a um hospital, ele poderia fazer um levantamento de informações em sua base de dados local, bem como estender esta busca em uma base de dados do hospital, de forma a tornar possível o acompanhamento da evolução do caso em que esteja trabalhando. Considerando que normalmente os pacientes não costumam portar os laudos anteriores e os exames que tenha feito anteriormente, esta ferramenta pode facilitar o acompanhamento da evolução deste caso.

Este cenário poderia ainda ser expandido para o processo de discussão com outros Médicos e especialistas que estejam em outros locais, tais como radiologistas, anatomopatologista, bioquímicos e outros. Mas neste cenário vislumbra-se apenas o enfoque regional, onde os atores se conhecem pessoalmente e atuando na mesma localidade, como por exemplo, os consultórios dos postos de saúde dos bairros ou cidades circunvizinhas e até mesmo os consultórios particulares, onde é feito normalmente, o primeiro atendimento e em seguida o encaminhamento do paciente.

Ou ainda em Clínicas que realizam os exames especializados para todo o estado como, por exemplo, a Clínica Blasinger, Benz & Buddenbrock na Alemanha que realiza a maioria dos exames radiológicos especializados para o estado Rheinland-Pfalz e possui a maioria dos laudos anteriores gravados em arquivos do tipo “.doc”.

1.2.3 Cenário #3 – Redes Médico-Hospitalares Compartilhando Informações Globalmente

Pode-se também expandir os cenários anteriores, levando os médicos a procurar e discutir casos com especialistas de outros hospitais, formando uma rede de médicos com a possibilidade de disponibilizar o conteúdo textual do sistema

legado sob a forma informações estruturadas para realizar buscas mais eficientes, intercambiar dados padronizados com centros de excelência em determinados tipos de tratamentos.

Este tipo de atividade pode acontecer não somente dentro de uma mesma região, estado ou país, mas poderia expandir-se para outros países com a utilização de nomenclaturas internacionais na área da saúde. Por exemplo, profissionais do Brasil poderiam ter removida a barreira do idioma para trocar informações confiáveis com profissionais que esteja na Alemanha, assim como de toda a Europa, América, África ou até mesmo em locais inóspitos como a Antártica.

Ou até mesmo, um determinado paciente que está recebendo atendimento em um local diferente dos locais onde já possui algum prontuário e laudos registros, no Brasil ou no Exterior, seja, pela troca de domicílio ou viagem de férias, negócios ou estudo, pela inexistência da especialidade requerida em sua cidade de origem ou qualquer outra razão. Mas se os locais onde estão estes dados não os possuir de forma padronizada, tornar-se muito difícil a interoperabilidade das informações entre os sistemas e os profissionais da área da medicina ficam sem importantes informações ou necessitam perguntar tudo novamente ao paciente que nem sempre se recorda de todos os detalhes, isso sem considerar condições onde ele nem mesmo esteja consciente para prestar as informações.

1.2.4 Cenário #4 – Ferramenta de Ensino Aprendizagem na Medicina.

Outro cenário possível de ser concebido são preceptores e residentes médicos poderem discutir e buscar casos semelhantes durante os cursos de residência médica, assim como nos dois cenários anteriores, bem como ainda poderia ser criado um banco de informações estruturadas e padronizadas de casos raros, devidamente anonimizados, por questões éticas, visando ao ensino presencial ou mesmo ferramentas de apoio ao EAD – Ensino a distância, para casos de cursos de atualização ou educação continuada.

Novamente, pode-se vencer as barreiras geográficas e lingüísticas, fazendo-se uso ou disponibilizando-se a base de informações para qualquer ponto do mundo, sem a necessidade de interpretes.

1.3 Funcionalidades

Visando aos cenários de aplicação, faz-se necessário criar um ambiente que trabalhe a descrição textual dos casos, interpretá-los computacionalmente e convertê-los para um padrão estruturado que permita melhor manipulação e troca de dados e informações com sistemas da área médica, visando a melhorar a prática dentro desta área, permitindo assim, adaptar-se e satisfazer a necessidade de características dos casos de difícil diagnóstico também em locais que possuam bases de dados em forma textual. Dentre as funcionalidades requeridas pelo sistema encontram-se:

Aceitar trabalhar com Laudos de mais de um idioma;

Gerenciamento de Nomenclatura internacional controlada multi-idiomas;

Gerenciamento de palavras conforme sua classe gramatical;

Recuperar e codificar a Informação contida nos laudos em padrão texto;

Estruturar as informações recuperadas e codificadas;

Gerar as informações estruturadas em um padrão internacional para intercâmbio de dados;

Exportar os dados para outros sistemas.

1.4 Objetivos.

O Objetivo geral é o desenvolvimento de uma metodologia de interpretação, recuperação e codificação inteligente de laudos médicos, independente de idioma,

visando o resgate de informações de casos antigos que se encontram sob a forma de textos livres em sistemas legados.

1.4.1 Objetivos específicos

Os objetivos específicos são os relacionados a seguir:

Proposição de métodos de que permitam dentro da linguagem médica, especificamente da Radiologia, interpretar computacionalmente laudos sob a forma de texto livre visando à conversão para padrões estruturados;

Desenvolvimento do protótipo da ferramenta baseada nos métodos propostos usando o processamento de linguagem para análise e interpretação de lados com casos controlados com diagnósticos previamente conhecidos para ajustes e aperfeiçoamento deste protótipo;

Teste do protótipo com casos antigos e novos de diagnósticos, apresentando-os para opinião de especialistas da área médica, também visando os ajustes e aperfeiçoamento do protótipo;

1.5 Estrutura do Trabalho.

O presente trabalho está compostos por 7 capítulos. No capitulo 2 estão os requisitos gerais e a descrição do processo de geração de laudos.

No capitulo 3, está o estado da arte em processamento de linguagem natural de forma geral e na área médica.

No capítulo 4, encontram-se as informações referentes aos padrões e nomenclatura internacionais como a CID-10 e SNOMED.

No capitulo 5, está a análise detalhada de requisitos do protótipo.

No capitulo 6, está o Modelo de análise de Laudos e alguns exemplos de análises de laudos realizados.

No capitulo 7, encontram-se as considerações finais e as recomendações para futuros trabalho.

2 Análise de Requisitos

Após entrevistas com médicos no Brasil da Clinica DMI e Curso Residência Médica em Cirurgia Geral do Hospital Regional Alto Vale e na Alemanha na Radiologische Gemeinshaftspraxis Buddenbrock, Blasinger und Benz, definiram os principais requisitos do sistema.

O requisito e foco central é permitir a recuperação de laudos médicos provenientes de sistemas legados em padrão texto, transformando-os em laudos estruturados em um padrão internacional para intercâmbio de informações na área da saúde. Para tanto, faz-se necessária a leitura de arquivos que se encontram em formato text (.txt) padrão ASCII exportados de antigas bases de dados, tal como Dataflex, Cobol e outros, que possuem padrões próprios para seus arquivos mas muitas vezes ainda se encontram em funcionamento, como por exemplo, no Hospital Regional Alto Vale. Também se faz necessário tratar arquivo do tipo documento (.doc) provenientes do processador de textos como o Microsoft Word e compatíveis, como por exemplo, a clinica Radiologische Gemeinshaftspraxis Buddenbrock, Blasinger und Benz da Alemanha.

Como requisito para o compartilhamento de informação entre os profissionais da saúde dos dois paises, do sistema, ele deve ser independente de idioma, permitindo que seja escolhido o idioma dos laudos em que se deseja trabalhar.

Baseado neste, o próximo requisito é o suporte à manutenção e/ou à importação/exportação do cadastro de nomenclaturas internacionais de termos na área da saúde, principalmente os termos médicos utilizados para gerar os laudos. Surgindo assim, a necessidade de permitir a classificação das palavras comumente usadas nos laudos formando um Thesaurus codificado compatível entre os dois países.

Outro requisito do protótipo de software a ser desenvolvido deve ser um módulo possível de ser integrado aos softwares da plataforma Cyclops de softwares para telemedicina e teleradiologia. Visando manter a compatibilidade, necessitará

ser desenvolvido na linguagem Smalltalk, que já é amplamente usada pelo grupo de pesquisa The Cyclops Project (http://cyclops.telemedicina.ufsc.br/) e que possui o programa de cooperação científica bilateral Brasil-Alemanha.

Um último requisito é a necessidade de utilizar programas softwares para o desenvolvimento do protótipo que possuam licenças de uso livres de cobrança, permitindo que o protótipo de software não possua custos de desenvolvimento e também para a sua futura utilização.

3 Estado da Arte

3.1 Processamento de Linguagem Natural.

O Processamento de Linguagem Natural (sigla em inglês NLP), segundo Peter Jackson [JACKSON2002] é o conjunto de métodos formais para analisar textos e gerar frases escritas em um idioma humano. Normalmente, computadores estão aptos a compreender instruções escritas em linguagens de computação como o Java, C, Cobol, Basic e outras, mas possuem muita dificuldade em entender comandos escritos em uma linguagem humana. Isso se deve ao fato de as linguagens de computação serem extremamente precisas, contendo regras fixas e estruturas lógicas bem definidas que permitem o computador saber exatamente como deve proceder a cada comando. Já em um idioma humano, uma simples frase pode conter ambigüidades, nuances e interpretações que dependem do contexto, do conhecimento do mundo, de regras gramaticais, culturais e de conceitos abstratos.

Um dos objetivos do Processamento de Linguagem Natural é fornecer aos computadores a capacidade de entender e compor textos. E "entender" um texto significa reconhecer o contexto, fazer as análises sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos e até aprender conceitos com os textos processados.

A história de Processamento de Linguagem Natural refere-se aos primeiros destaques importantes da história da IA - Inteligência Artificial ou Aplicada, vão de Turing a Minsky, de Holland a Schank. O primeiro estágio na história de IA inicia-se com McCulloch e Pitts (1943) e é conhecido como a era do "Olhe, mamãe, sem as mãos!" (Look, Ma, no hands era), [RUSSEL & NORVIG1995]. Um segundo período pode ser caracterizado pela busca de Newell e Simon pelo General Problem Solver (GPS), um resolvedor de problemas genéricos (1961).

A Inteligência Aplicada envolve o desenvolvimento de sistemas de computadores inteligentes que tratam de problemas em microworlds (microdomínios). Definido por Minsky em [RUSSEL & NORVIG1995], microworlds são domínios limitados de aplicação. Esta etapa mais recente também é caracterizada pela busca da técnica mais apropriada para resolver cada tipo de problema. Esta pesquisa é uma instância nessa busca. Existe o confrontamento com o problema de como desenvolver sistemas capazes de resolver problemas complexos compostos por distintas tarefas. Na medida em que deve haver uma técnica mais apropriada para resolver cada tarefa, culmina hoje a era dos agentes inteligentes, aonde várias ferramentas de IA e de outras áreas do conhecimento são combinadas em um único sistema inteligente que administra forças para resolver tarefas.

A IA também é o campo de estudo destinado ao desenvolvimento de sistemas computacionais que simulam algum aspecto de cognição humana. São discernidas duas orientações a partir desta definição; a tecnológica, associada ao desenvolvimento de programas inteligentes; e o científico, que trata dos aspectos teóricos de cognição humana.

Especificamente, dentro do Processamento de Linguagem Natural destaca-se inicialmente o Sistema Eliza trabalho de Joseph Weizenbaum da década de 60, que se assemelhava a um psiquiatra, fazendo a análise de perguntas e respostas feitas ao usuário e aprendendo com isto.

Nessa mesma linha, encontra-se o Sistema RAPIER da Universidade do Texas em Austin, que combina o aprendizado do Eliza com técnicas de programação de lógica indutiva, permitindo seu uso em diversos domínios, teve casos de sucesso na busca de E-mails contra o terrorismo na América Latina e também informações em documentos empresariais.

O processamento de Linguagem Natural pode ser dividido, segundo [WANGENHEIM1993], em níveis ou categorias conforme as Figuras 1 e 2.

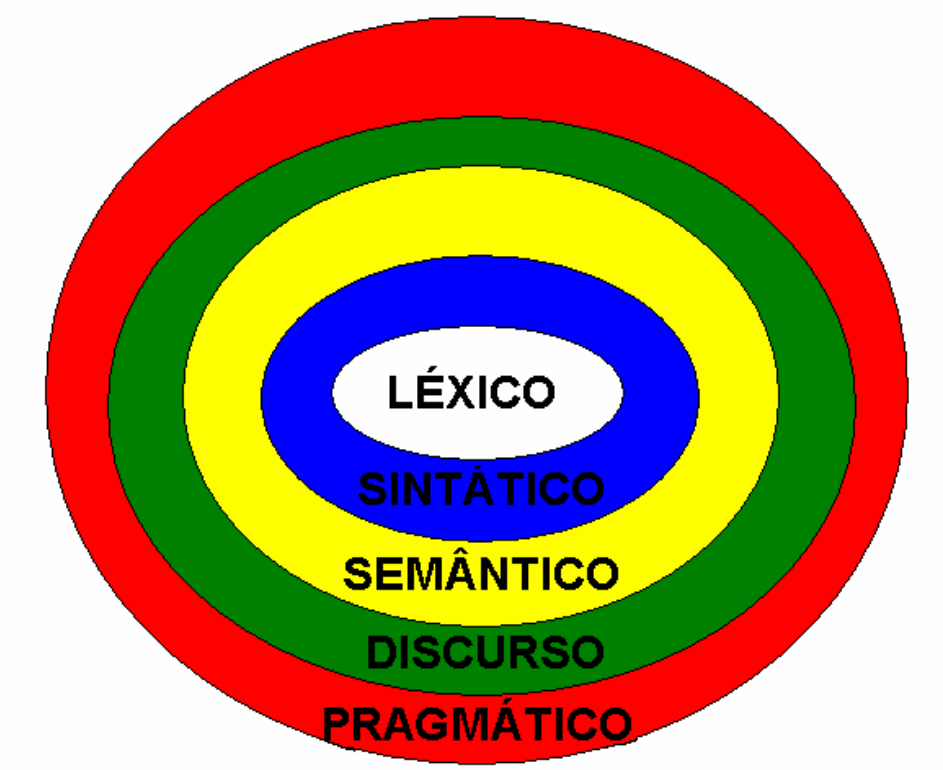


Figura 1 – Níveis de Processamento de Linguagem Natural

Fonte: [WANGENHEIM1993]

No nível Léxico, verifica-se a definição das palavras e sua classe, assim como a correta escrita. No Sintático, analisa-se a formação e estrutura das frases e sentenças.

Na análise semântica, faz-se a verificação do significado associado à estrutura Sentencial, na análise de discurso, faz-se a verificação do domínio intersentencial, ou seja, contexto e a interferência que as frases provocam entre si para a compreensão do texto.

A análise pragmática engloba o contexto lingüístico e experiências na área do texto a sofrer o processamento. A comparação entre estes diversos níveis de processamento está na Figura 2.

Níveis de Processamento de Linguagem Natural	
Nível	Abordagem
Léxico	Definição das Palavras e sua Classe
Sintático	Formação e estrutura de Frases e Sentenças
Semântica	Significado associado à estrutura Sentencial
Discurso	Domínio Intersentencial = Contexto.
Pragmático	Contexto Lingüístico + Experiências Humanas,

Figura 2 - Significado Níveis de Processamento de Linguagem Natural

Fonte: [WANGENHEIM1993]

3.2 Processamento de Linguagem Natural na área médica.

Dentre os trabalhos publicados sobre o processamento de linguagem natural para área médica, cita-se o trabalho de Robert E. Williamson, na Biblioteca Nacional de Medicina (BNM) [WILLIAMSON1985] referente ao Annod, um sistema que implementa um conjunto de técnicas lingüísticas e empíricas que permitem a recuperação de informações de linguagem natural em resposta a perguntas de linguagem natural. O sistema é baseado no sistema de recuperação de documentos do Dr. Gerard Salton (SMART) fazendo parte de um sistema interativo de gerenciamento de texto, SIGT. A experiência com recuperação de informações da base de conhecimentos de hepatite da BNM apresenta um arquivo texto hierárquico. As técnicas usadas no Annod incluem: remoção de palavras, marcação automática de palavras, um algoritmo complexo de combinação empírica de medida de similaridade e técnicas desenvolvidas para permitir rápidas respostas. Os testes demonstraram alta eficiência na identificação das porções de um texto que seja relevante aos usuários, na busca de informações sobre hepatite.

No artigo intitulado “Ontologia guiada para descoberta de conhecimento em base de dados”, Joseph Phillips e Bruce Buchanan [PHILLIPS E BUCHANAN2001]

apresentam um trabalho referente uma a nova metodologia para aumentar os conteúdos semânticos de ontologia para guiar a descoberta de conhecimentos em base de dados. O sistema analisa através de novas bases de dados, para obter informações sobre os tipos de variáveis que o usuário verifica. Então usa-se essa informação no contexto de uma ontologia compartilhada, para guiar, de forma inteligente, os processos combinatórios em potencial de uma construção de apresentação. O sistema também aprende a cada vez em que é utilizado, facilitando a verificação de tarefas do usuário nas utilizações subseqüentes. Neste processo, a descoberta de conhecimento em base de dados (DCBD) é o processo iterativo que está implícito em uma grande e diversa base de dados.

Ou seja, um processo iterativo de herança e hereditariedade de seleção de dados, processamento, transformação em uma forma trabalhável, uso de data mining e interpretação dos resultados. O processo é intensivo em conhecimento, porque todos os passos requerem conhecimento domínio-específico, para decidir quais operações podem se mostrar mais úteis, vindas de um grande conjunto em potencial. A escolha de vocabulário também é crucial para programas de descoberta. A crescente importância de ontologias compartilhadas em larga escala é demonstrada em vários projetos para construí-los em uma variedade de domínios. Dois exemplos são Ontolíngua, e o Sistema Unificado de Linguagem Médica. As Ontologias permitem racionalização em domínios através da codificação de conhecimento específico.

[ABID e MANICKAM2002] em “Extracting Case Structures from XML-Based Electronic Patient Records” empregam uma abordagem automatizada para gerar casos médicos em um sistema de raciocínio baseado em casos (CBR). Considerando que sistemas de diagnósticos médicos vêm a exigir um volume crítico de diagnósticos de qualidade atualizados, para que estes casos descrevam uma metodologia de resolução de problemas médicos. Em termos práticos, consideram desafiante requerer dos peritos em medicina, mapeamento de sua experiência em conhecimento para uma formalização computacional que não lhes é familiar. Eles propuseram uma abordagem de aquisição de conhecimento médico que gera registros médicos eletrônicos (RMEs) como uma fonte alternada para CBR-

compliant-cases. Uma metodologia multi-estágios apresenta: (a) coleção de RME heterogêneas vindas de repositórios acessíveis via Internet de RME através de agentes inteligentes; (b) transformação automatizada de ambas as estruturas e conteúdos genéricos de RME para casos especializados CBR-compliant, e (c) estima e indução do peso de cada um dos casos e atributos definidos.

[BAKKEN e WARREM2002] trabalharam na avaliação da utilidade de dois modelos de terminologia para integração de conceitos diagnósticos para termos clínicos do SNOMED (SNOMED-CT). Como métodos, primeiramente dissecaram as frases com termos de diagnósticos de duas fontes de terminologias NANDA e Omaha System para as categorias de semântica do CEP - Comitê Europeu para Padronização da estrutura categorias e o modelo ISO de referência de terminologia. Depois, analisaram criticamente as semelhanças entre as ligações semânticas nos modelos CEP e ISO e as ligações semânticas usadas para definir formalmente em SNOMED-CT. Resultando na demonstração de que o foco, portador, assunto de informação e julgamento estavam presentes em 100% dos termos no Omaha e NANDA das frases. Os termos das frases Omaha não continham descritores além dessas consideradas imperativas nos modelos CEP e ISO. A comparação entre as ligações semânticas precisava modelar as duas fontes de terminologias para a integração. Em conclusão, dos autores, os resultados sustentam a utilidade potencial dos modelos CEP e ISO para integração com o SNOMED-CT.

No trabalho de Peter L. Elkin e outros [ELKIN e BROWN 2002] “Guideline and quality indicators for development, purchase and use of controlled health vocabularies”, constata que tanto o desenvolvedor quanto o usuário de terminologias de saúde controladas, requerem mecanismos para sistemas de comparação de terminologias. Pelo controle de vocabulários de saúde, referem-se às terminologias e sistemas terminológicos feitos para representar dados clínicos em uma consistente com a prática de entrega de cuidados clínicos de hoje em dia. Critérios compreensíveis para a avaliação de tais sistemas, historicamente têm tido falhas e o critério conhecido é aplicado inconsistentemente. Segundo os autores, apesar de existirem vários trabalhos que descrevem especificamente as características desejadas de um vocabulário de saúde controlado, não há um guia consistente para

avaliações de terminologias de referência, o que os ajudaria a comparar sistemas terminológicos para se referenciar, o que os ajudaria a comparar implementações de sistemas terminológicos de igual para igual. Eles propuseram um guia que serve para preencher as falhas entre enumerações acadêmicas de características de terminologias desejáveis e a implementação prática ou avaliação rigorosa que rendem dados comparáveis, considerando a qualidade de um ou mais vocabulários de saúde controlados.

No trabalho “Data quality probes exploiting and improving the quality of electronic patient record data and patient care” de Philip J.B.Brown e Victoria Warmington [BROW e WARMINGTON2002] destacam que uma confiança cada vez maior está sendo colocada em registros médicos eletrônicos para apoiar cuidados clínicos e alcançar o aprimoramento dos padrões de qualidade. Para que os sistemas de informações clínicos entreguem uma excelência em dados, precisam ser completos, consistentes e precisos. Essa captura de dados é crítica, mas forma apenas uma parte no procedimento de entregar uma informação de qualidade no cuidado com pacientes durante o encontro clínico-paciente. Vários processos estão envolvidos nesse encontro, cada qual deve ser feito sem falha alguma para entregar um resultado perfeito. O trabalho aponta um método de acessar a qualidade desses processos envolvidos na provisão de cuidados clínicos e qualidade de dados em um sistema de informações clínicos. Propõe o princípio de Prova de Qualidade de Dados (PQD) para assessorar o desempenho de todo o sistema integrado. A sua característica principal e que gera uma “Consulta” na qual as predições de conhecimento clínico não devem recuperar nenhum caso num sistema, operando com absoluta ausência de falhas. Qualquer caso recuperado (que falha o PQD) indica um erro tanto na qualidade do dado como em julgamentos clínicos. Essa abordagem é aplicada praticamente com o paradigma de uma família do Reino Unido testando a hipótese de que uma série de PQDs podem fornecer um método valioso para monitoração tanto da precisão dos dados como dos Sistemas Clínicos de informação e da provisão de qualidade no cuidado com o paciente.

Riccardo Bellazi Stefania Montaini em seu artigo, [BELLAZI e MONTAINI 2001] “Cased-Based Reasoning for medical knowledge-based systems” afirmam que

no domínio médico, tipos diferentes de conhecimento são tipicamente encontrados. Conhecimento operacional, adquirido durante a prática de cada dia, e relatando habilidades de especialistas, é armazenado no sistema de informação do hospital (HIS). Por outro lado, bem assessorados, formalizados conhecimentos médicos são relatados em livros texto e guias clínicos. A proposta feita por eles é que todas essas informações heterogêneas deveriam ser distribuídas de forma segura, e tornar-se disponível para médicos na forma e no tempo correto, para poder apoiar a tomada de decisão. Sob o ponto de vista dos autores, entretanto, um sistema para apoiar a tomada de decisão não pode ser concebido como uma ferramenta independente, hábil para substituir a sabedoria humana, mas deve ser integrada com a tarefa de gerenciar o conhecimento. Do ponto de vista metodológico, CBR (case-based reasoning) provou que é um paradigma adequado de raciocínio para gerenciar conhecimento do tipo operacional. Por outro lado, raciocínio baseado em regras é historicamente uma das abordagens mais bem sucedidas para abordar o conhecimento formalizado. Para ter a vantagem em todos os tipos de conhecimento, eles propõem uma metodologia de raciocínio multi modal, que integra raciocínio baseado em casos e raciocínio baseado em regras, para apoiar a detecção de contexto, recuperação de informação e apoio à decisão. Segundo os autores, esta metodologia tem sido testada, com sucesso, na aplicação de gerenciamento de pacientes diabéticos.

No trabalho “A Delphi technique as a method for selecting the content of an electronic patient record for asthma.” [STEENKISTE E JACOBS2002], os autores consideram que um registro eletrônico de paciente com dados de doenças específicas, pode suportar aprimoração da qualidade dos cuidados de pacientes com doenças crônicas. A estrutura e conteúdo de registro podem somente ser acessados por clínicos em cooperação com especialistas de Tecnologia da Informação (TI), porque, o resultado terá uma relevância clínica de fácil acesso, e ajustáveis às informações necessárias em diferentes trabalhadores de cuidados primários. Eles aplicaram um procedimento modificado de delphi, um método caracterizado pela escrita anônima de um “painel especialista”. O painel tem que concordar com a questão se ela deve ou não ser incluída no registro eletrônico de paciente. As questões, para os comentários escritos, foram preparadas por um

comitê de decisão composto por clínicos gerais, cientistas da área de saúde, especialistas em asma, gerenciadores de tratamento de doenças ou especialistas de TI, baseado nas instruções para diagnósticos e tratamento de asma do Dutch College of General Practitioners (DCGP). Quando há o acordo com o painel de até 70%, é enviado um formato modificado para o “painel especialista” para uma reavaliação. Os autores conseguiram como resultados, após três rodadas de comentários escritos com 95 itens potenciais sendo discutidos com o “painel especialista”. Na primeira rodada, eles selecionaram 50 itens relacionados ao diagnóstico de asma e 22 relacionados com o tratamento de asma. Durante a segunda rodada, 17 itens ainda estavam em discussão e 6 rejeitados. Nas rodadas subseqüentes, o “painel especialista” acessou o melhor formato de registro.

Segundo [LIU1996], os vocabulários controlados têm sido usados para fins de unificação e evitar distorções terminológicas encontrados em vários campos de aplicações. Essa unificação leva a uma melhor administração de informações e aperfeiçoa a comunicação entre várias partes.

A maioria dos trabalhos citados tem abordagem na busca de padrões ou ferramentas para exploração conhecimento, mas em comum, destacam a utilização de terminologias ou nomenclatura da área da medicina, tais como, o SNOMED. Citam problemas com a qualidade das informações como fatores principais de casos de sucesso das aplicações ou metodologias propostas. Outro fator que merece destaque é a carência de trabalhos voltados a aplicações com a Língua Portuguesa, visto que a maioria das terminologias encontra-se prioritariamente em Inglês, possuindo, somente em alguns casos, versões para outros idiomas como o Alemão, Francês e Espanhol. Mesmo sendo o Português a oitava língua mais falada do planeta (terceira entre as línguas ocidentais, após o Inglês e o Castelhano) [MEDEIROS1998]. Outro fato interessante é que somente em 1996 foi criada, em Lisboa, a “Comunidade dos Países de Língua Portuguesa (CPLP)”. Entidade com a finalidade de reunir os sete países lusófonos existentes - Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, Portugal e São Tomé e Príncipe para cooperação na área da Saúde e Educação [MRE1996].

4 Nomenclaturas Internacionais e Padrões de informações médicas.

A padronização no armazenamento e no uso do vocabulário médico é fundamental para reunir informação clínica no cuidado ao paciente, para recuperar informação no manejo da doença ou para pesquisa, assim como para conduzir a análise de resultados.

4.1 CID-10.

A CID-10 - Classificação Internacional de Doenças é periodicamente revisada e publicada pela OMS - Organização Mundial da Saúde, e a mesma contém capítulos referentes a diversos grupos de doenças.

A CID-10 constitui-se de uma família de documentos, que se presta à utilização para uso clínico, educacional e assistencial em geral. Outros documentos que compõem esta família, ainda em desenvolvimento ou não publicados, compreendem critérios diagnósticos para pesquisa, uma classificação multiaxial, outra específica para serviços de cuidados primários e, finalmente, um glossário.

No Brasil, com base no compromisso assumido pelo Governo Federal, quando da realização da 43ª Assembléia Mundial de Saúde, o Ministério da Saúde, por intermédio da portaria nº 1.311, de 12 de setembro de 1997, definiu a implantação da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - CID-10, a partir da competência de janeiro de 1998, em todo o território nacional. Segundo esta portaria do Ministério da Saúde, a CID-10 está em vigor no Brasil e todo e qualquer diagnóstico deverá ser relatado segundo as normas constantes na CID-10, tarefa que se revelará mais do que um procedimento burocrático a ser seguido, pois servem para as estatísticas no âmbito nacional e internacional na área da saúde.

Os documentos CID-10 fornecem em suas Descrições Clínicas e Diretrizes Diagnósticas, critérios específicos para que um determinado diagnóstico possa ser estabelecido. Caso o quadro clínico do paciente preencha parcialmente os critérios

requeridos para aquela condição, por exemplo, um número menor de sintomas ou uma duração menor na apresentação do quadro, um diagnóstico provisório ou tentativo poderá ser atribuído a ele.

.

Mas como contempla prioritariamente, o diagnóstico não contém a maioria dos demais itens do vocabulário médico, ficando assim incompleta para ser à base da análise dos laudos que descrevem os achados e resultados dos exames. A lista da divisão das categorias conforme os capítulos da CID-10 divulgadas pelo CBCD e DATASUS esta no Quadro 2.

Quadro 1: CID-10: Categorias CBCD e DATASUS

CID-10 Lista de categorias de três caracteres		
Capitulo	Códigos	Conteudo
I	(A00-B99)	Algumas doenças infecciosas e parasitárias
II	(C00-D48)	Neoplasias [tumores]
III	(D50-D89)	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
IV	(E00-E90)	Doenças endócrinas, nutricionais e metabólicas
V	(F00-F99)	Transtornos mentais e comportamentais
VI	(G00-G99)	Doenças do sistema nervoso
VII	(H00-H59)	Doenças do olho e anexos
VIII	(H60-H95)	Doenças do ouvido e da apófise mastóide
IX	(I00-I99)	Doenças do aparelho circulatório
X	(J00-J99)	Doenças do aparelho respiratório
XI	(K00-K93)	Doenças do aparelho digestivo
XII	(L00-L99)	Doenças da pele e do tecido subcutâneo
XIII	(M00-M99)	Doenças do sistema osteomuscular e do tecido conjuntivo
XIV	(N00-N99)	Doenças do aparelho geniturinário
XV	(O00-O99)	Gravidez, parto e puerpério
XVI	(P00-P96)	Algumas afecções originadas no período perinatal
XVII	(Q00-Q99)	Malformações congênitas, deformidades e anomalias cromossômicas
XVIII	(R00-R99)	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte
XIX	(S00-T98)	Lesões, envenenamento e algumas outras conseqüências de causas externas
XX	(V01-Y98)	Causas externas de morbidade e de mortalidade
XXI	(Z00-Z99)	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde

Fonte: DATASUS

A CID-10 adota, na identificação dos diversos quadros, um código alfa-numérico composto por uma letra e até quatro caracteres numéricos. Cada capítulo da CID-10 é identificado por uma letra, por exemplo, o Capítulo V identificado pela

letra F. Ou seja, toda vez que um código da CID-10 se inicie pela letra F, aquela categoria diagnóstica identifica um transtorno mental ou de comportamento, outros exemplos são as doenças infecciosas intestinais (A00 – A09). O código básico inclui, além desta letra, outros dois caracteres numéricos, sendo que a combinação destes três, uma letra e dois números, permitem 100 categorias diagnósticas principais

4.2 Nomenclatura SNOMED

A SNOMED “The Systematized Nomenclature of Human Medicine” é uma das mais completas nomenclaturas multiaxiais criadas para indexar o conjunto de registros médicos, possuindo tradução em diversos idiomas, tais como, Inglês Alemão e Espanhol.

A SNOMED internacional foi formada em setembro de 1993, mas já vinha sendo traçada desde o início dos anos 60, com a denominação de Systematized Nomenclature for Pathology (SNOP). Ela inclui sinais e sintomas, diagnósticos e procedimentos; e seu projeto único irá permitir a integração completa de todas as informações médicas, em um registro médico eletrônico dentro de uma estrutura única de dados. A composição da nomenclatura SNOMED possui alguns eixos para distinção dos termos e facilidade de localização dos mesmos

O eixo topográfico ou anatômico tem seus códigos iniciados com a letra T indicando anatomia funcional para medicina humana e veterinária. No âmbito Morfológico com a letra M, compreende as alterações encontradas nas células, tecidos e organismos. Os Diagnósticos indicados pela letra D indicam a classificação de condições reconhecidas clinicamente, encontradas na medicina humana e veterinária.

Já os Procedimentos indicados pela letra P com procedimentos administrativos, diagnósticos e terapêuticos, enquanto o Funcional - F: sinais e sintomas; fisiologia e fisiopatologia dos processos da doença e Organismos vivos - L: organismos com vida de etiologia significante na doença humana e animal.

Existem ainda o eixo Químico - C: drogas, produtos biológicos e manufaturados farmacêuticos, Agentes físicos, ações e forças - A: compêndio de

ações físicas, perigos físicos, e forças da natureza, Contexto social - S: condições sociais e suas relações de importância para a medicina., Ocupações - J: termos que descrevem a ocupação e por último de uso Geral - G: ligações, descrições e qualificações que se associam ou que modificam os termos contidos em cada eixo.

A Figura 3 apresenta um modelo dos principais tipos de termos dentro da nomenclatura SNOMED, onde tem se uma idéia da abrangência e importância desta nomenclatura, motivos pelos quais ela tem se difundido rapidamente nos meios médicos internacionais. Aliado a isto, também o comitê organizador do SNOMED buscando sempre atualizá-la nos mais diversos idiomas tem buscando parceira com médicos nas mais diversas especialidades dentro de seus países de atuação.

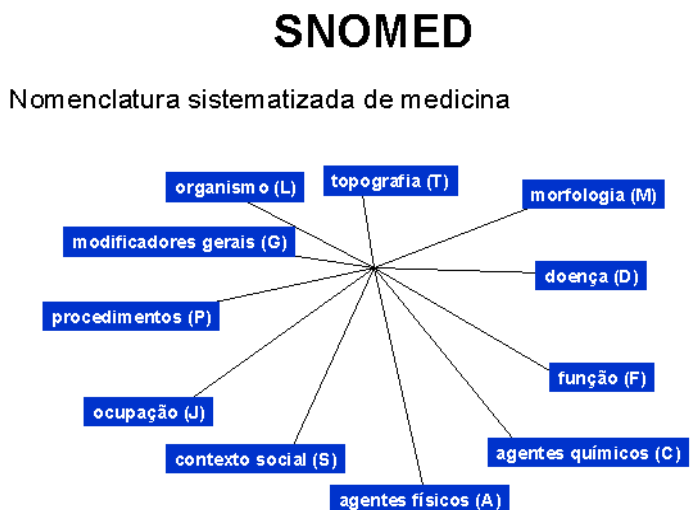


Figura 3 –Modelo dos Eixos da Nomenclatura SNOMED.

Fonte: SNOMED

A sua Sistematização compreende uma combinação de alguns destes eixos. Por exemplo, um diagnóstico completo na SNOMED consiste em um código topográfico, um código morfológico, um código de organismo vivo e um código funcional. um diagnóstico, bem definido, é estabelecido pela combinação destes quatro códigos existe. Por exemplo, a doença com o código D-13510 Pneumonia pneumocócica é equivalente à combinação de: T-28000 (código topográfico para

pulmão); M-40000 (código morfológico para inflamação) e L-25116 (código para Streptococcus pneumoniae do eixo de organismos vivos).

Um outro exemplo é a doença dermatite atópica com o código D-10130 é composta pelos seguintes eixos: T-01000 (código topográfico para pele), M-4300 (código morfológico para inflamação crônica), M-01735 (código morfológico para eritema papulovesicular), F-C3000 (código funcional para reação de hipersensibilidade alérgica) e F-A2300 (código funcional para o sintoma coceira). A Quadro 3 apresenta parte da organização da SNOMED extraída do livro em alemão [WIGERT 1984] entre as páginas 51 a 57 referente a alguns termos chave do aparelho respiratório.

O que pode ser notado é que o termo principal em cada um dos casos está escrito próximo ao termo em Latin e possui uma lista de termos sinônimos no idioma onde a SNOMED está traduzida. Um exemplo é o pulmão em Português que tem origem no Latin Pulmone conforme o Dicionário Universal da Língua Portuguesa, já no idioma Alemão é indicada a mesma origem só que apenas Pulmo conforme o Wörtebuch de Medizin [ZETKIN 1980].

O comitê do SNOMED [SNOMED2004] está sempre preocupado com a constante evolução e revisão dos termos, aplicados em conjunto às tecnologias existentes e combinadas com novas, para criar as aplicações que venham a dar suporte a esta prática médica. A evolução da SNOMED pode ser visualizada na Quadro 2:

Destaca-se também o trabalho do comitê em difundir e ampliar os conceitos e termos médicos, procurando parcerias em todos os países para a tradução e emprego da nomenclatura de forma universal.

Quadro 2: Exemplo da Estrutura Snomed

Exemplo da Estrutura do SNOMED			
Alemão			Português
Código	Termo Principal	Sinônimos	Tradução
T20000	Apparatus respiratorius	Atemwege	Aparelho Respitatório
		Atumengsorgane	
		Luftwege	
		Respirationstrakt	
T21000	Nasus	Nas	Nariz
		Nase	
		rhin	
T22000	Sinus paranasalis	Nasennebenhöhle	Seios Paranasais
		Sinus nasi acessorius	
T2500	Trachea	Trachea	Traquéia
		Luftröhre	
		Trache	
T26000	Bronchus	Bronchus (principalis)	Brônquios
		Bronch	
		Hauptbronchus	
		Stammbronchus	
T27000	Bronchiolus	bronchiol	bronquíolos
		Bronchiole	
		Bronchulus	
T28000	Pulmo	Lunge	Pulmão
		pneumon	
		pulmom	
T29000	Pleura	Brustfell	Pleura
		pleur	

Fonte: [WIGERT 1984]

Além das vantagens que os hospitais obtém com o uso do SNOMED, como nomeclatura padronizada vocabulário unificado e facilidade de uso de sistemas informatizados, elas também são revertidas aos pacientes e demais profissionais da área da Saúde passam a obter ou usufruir de atendimento com maior qualidade, como por exemplo facilidade de comunicação de seu caso com profissionais de outros países.

Quadro 3: Histórico da Nomenclatura SNOMED.

SNOMED History		
Yaer	Name	Terms
1965	Systematized Nomenclature of Pathology (SNOP)	-
1976	Systematized Nomenclature of Medicine (SNOMED)	44.587
1984	Snomed em Alemão	+50.000
1986	SNOMED	+50.000
1993	SNOMED Internacional	130.580
1997	SNOMED Internacional ver 3.5	156.602
1998	SNOMED RT – A Reference Terminology	300.000
2002	SNOMED CT – Clinical Terms®	913.000
2003	SNOMED CT – Clinical Terms®	1.3 milhões

Fonte: SNOMED [SNOMED2004]

4.3 HL7 - Health Level Seven

O padrão HL7 - Health Level Seven é um padrão ANSI aprovado pela American National Standard para troca de documentos eletrônicos na área da saúde. Ele permite que diferentes aplicações computacionais troquem conjuntos relevantes de informações médicas, clínicas e administrativas.

O HL7 é composto por formatos padronizados – os protocolos HL7, que especificam a implementação de interfaces entre diferentes aplicações computacionais. Estes protocolos proporcionam a flexibilidade necessária para permitir a compatibilidade de conjuntos de dados distintos, que apresentam necessidades específicas.

Ele engloba dados de admissão e alta de pacientes, observação clínica, solicitação de exame, resultado e cobrança, além de outros dados de sistemas médicos. Em conformidade com o Sistema Legal Brasileiro, todos os usuários e produtores de sistemas de informação de saúde, devem produzir um conjunto mínimo requerido de informações.

Diversos produtores de softwares para área da saúde vêm aplicando o padrão HL7 em suas ferramentas visando à integração com os diversos sistemas existentes internamente aos hospitais, assim como as estruturas governamentais na área da saúde. Um exemplo de aplicação utilizando o HL7 está na Figura 4, o sistema HRR - Health Resources Register System desenvolvido pelo Erickson para o registro de recursos na área de saúde.

O padrão HL7 apresenta suporte a comunicações para diversas linguagens de programação, sistemas operacionais e ambientes de comunicação, abrangendo desde aplicações Internet até sistemas integrados. Outros padrões que normalmente são usados juntamente ao HL7 são PACS e DICOM.

O padrão PACS - Picture Archiving and Communication Systems é um sistema de arquivamento e comunicação voltado para os diagnósticos por imagem que permite o pronto acesso, em qualquer setor, de imagens médicas em formato digital. O sistema PACS em conjunto com os sistemas de informação radiológica (RIS) e de informação hospitalar (HIS) formam a base para um serviço de radiologia informatizado. Isso se refere a um hospital, com um ambiente de rede amplo e integrado, no qual o filme foi completamente, ou em grande parte substituído por sistemas eletrônicos que adquirem, arquivam, disponibilizam e exibem imagens. A implantação de um serviço de radiologia informatizado traz melhorias no que se refere à acessibilidade e integração de informações, pela vinculação de imagens ao

registro médico eletrônico do paciente, e no que se refere à aplicação de novas técnicas e desenvolvimentos na aquisição, exibição e processamento de imagens.

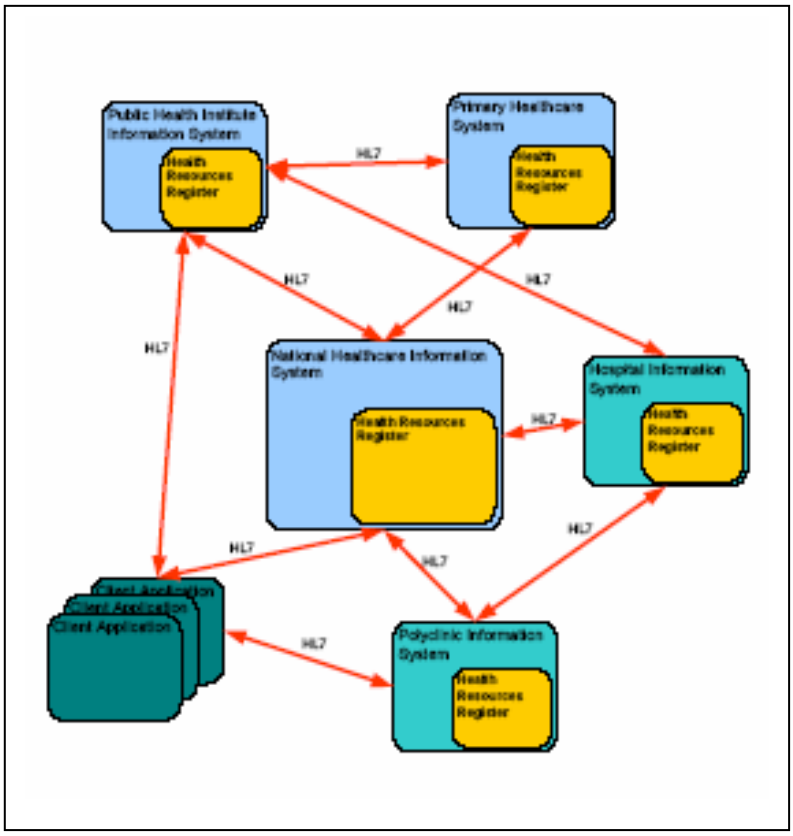


Figura 4 Modelo aplicação com o padrão HL7 da Erickson. (HRR - Health Resources Register System)

Fonte: HL7 [HL72004]

4.4 DICOM e DICOM Structured Report

DICOM (Digital Imaging Communications in Medicice) é um padrão que foi criado com a finalidade de se padronizar as imagens diagnósticas e, como Tomografias, Ressonâncias Magnéticas, Radiografias, Ultrassonografias, e outras. O padrão DICOM é composto por uma série de regras que permite que imagens médicas e informações associadas sejam trocadas entre equipamentos de imagem, computadores e hospitais. O padrão estabelece uma linguagem comum entre os equipamentos de marcas diferentes, que geralmente não são compatíveis, e entre

equipamentos de imagem e computadores, estejam esses em hospitais, clínicas ou laboratórios.

Já o padrão DICOM SR define como devem ser constituídos objetos de informação que codificam dados a respeito de exames, diagnóstico e tratamento, além de informações de contexto, como procedimentos que devem ser executados para o sucesso de um tratamento, e dados sobre profissionais de saúde envolvidos. Um objeto no padrão pode conter referências embutidas a imagens, eletrocardiogramas, e arquivos de áudio, bem como a outros documentos no mesmo padrão, e não contém especificações de como o laudo representado deve ser apresentado, ou impresso. Além disso, objetos no padrão utilizam terminologia controlada, como forma de evitar as ambigüidades das linguagens naturais, facilitar o entendimento automatizado do conteúdo, a busca por informações específicas e a internacionalização do conteúdo.

O termo relatório estruturado (structured report) tem significado diferente para pessoas diferentes. Um radiologista, pensando em termos de relatórios convencionais, pode imaginar um documento que consiste de uma hierarquia de títulos que contêm blocos de texto, talvez com alguns códigos ou senhas ao fim, para resumir os achados. O autor de um software, para fazer medidas obstetrícias por ultra-som, pode visualizar uma hierarquia aninhada de medidas numéricas relacionadas (como tamanho do fêmur do feto, cada um identificado com códigos individuais, e posteriormente agregados para prover medidas mais gerais, como médias ou estimativas de idade fetal). Da perspectiva de DICOM Structured Reporting (SR), o que unificam estas diferentes visões são:

- a presença de listas e relacionamentos hierárquicos;
- uso de conteúdo numérico ou codificado em adição a um texto simples;
- uso de relacionamentos entre conceitos;
- a presença de referências embutidas para imagens e objetos semelhantes.

No entanto, um relatório estruturado (ou de forma geral, um “documento estruturado”) é definido, mais pela maneira como ele é construído, do que pelo que ele contém. A palavra “relatório” (report) é realmente imprópria, já que DICOM SR pode transmitir (conter) qualquer tipo de conteúdo estruturado, não apenas relatórios. Documentos SR podem ser usados onde houver:

- necessidade de listas ou conteúdo estruturado hierarquicamente;
- necessidade de conceitos codificados ou valores numéricos;
- necessidade de referências a imagens, waveforms ou outros objetos compostos.

Eles não precisam ser documentos complexos, nem serem sempre de fácil compreensão por um ser humano.

O padrão DICOM SR estabelece como devem ser formados objetos compostos de informação que codificam dados a respeito de exames, diagnósticos e, tratamentos, além de informações de contexto, tais como procedimentos que devem ser executados para o sucesso de um tratamento, e dados sobre profissionais de saúde envolvidos.

Um objeto no padrão pode conter referências embutidas a imagens, eletrocardiogramas, e arquivos de áudio, bem como a outros documentos no mesmo padrão. Desta forma, um único objeto DICOM pode conter todas as informações referentes a um determinado tratamento.

Cada objeto codifica apenas informações semânticas, e não contém informações sobre como o documento representado pelo objeto deve ser apresentado, ou impresso. Portanto, cada implementação de prontuário eletrônico pode ter um formato para apresentação que lhe for mais adequado. Além disso, objetos no padrão fazem uso de terminologia controlada, o que evita as ambigüidades da linguagem natural, facilita o entendimento automatizado do conteúdo, a busca por informações específicas, e a internacionalização do conteúdo.

No modelo de dados do DICOM, um paciente tem um ou mais estudos, um estudo corresponde a uma visita do paciente a uma instituição de saúde. Cada estudo contém uma ou mais séries. Séries são seqüências de imagens ou de cortes de imagens. No caso de séries de documentos SR, isso não tem muita semântica. Uma série contém as imagens laudos DICOM, curvas e outros objetos DICOM.

O padrão DICOM SR define três diferentes classes SOP (do inglês Service Object Pair - par serviço-objeto) de laudos. Estas classes SOP são, em ordem crescente de complexidade e abrangência: Basic Text SR (Laudo Estruturado de texto), Enhanced SR (Laudo Estruturado aperfeiçoado), Comprehensive SR (Laudo Estruturado Abrangente). As diferenças entre estas classes são restrições impostas à estrutura do documento.

Em cada definição de classe SOP um IOD (do inglês Information Object Definition - Definição de Objeto de Informação) é combinado com um serviço de armazenamento. Um IOD é um modelo abstrato de dados orientado a objeto usado para especificar informações de objetos do mundo real [DELLANI2001].

O IOD Basic Text Structured Report é para relatórios com uso mínimo de códigos, tipicamente usados no título do documento e subtítulos e uma árvore hierárquica de subtítulos sob a qual podem aparecer textos e subtítulos. Referências a instâncias SOP (como imagens, formas de onda e outros documentos SR) são restritas às folhas (nodos que não possuem filhos) da hierarquia. Esta estrutura simplifica a codificação de documentos de texto como documentos SR, bem como sua renderização e apresentação.

O IOD Enhanced Structured Report é um superconjunto do Basic Text IOD. Também foi projetado para representar relatórios com uso mínimo de códigos e em adição ao Basic Text IOD, permite o uso de medidas numéricas com códigos para os nomes de medidas e unidades. Além disso, permite que referências a imagens ou formas de onda sejam acompanhadas de itens que identificam regiões de interesse espaciais e temporais.

O IOD Comprehensive é um superconjunto das classes Basic Text SR e Enhanced SR e especifica uma classe de documentos cujo conteúdo pode incluir

uma variedade de tipos de informação incluindo texto, medidas numéricas, referências a outras instâncias DICOM e regiões de interesse selecionadas destas instâncias. Um documento desta classe também permite relacionamentos ditos "por referência" entre os itens.

As informações em um Structured Report (SR) são agrupadas em 9 módulos cujos itens de informação se relacionam. Existe um módulo para informações sobre o paciente, como data de nascimento e peso, um módulo para informações gerais a respeito do documento, como por exemplo, nomes de pessoas responsáveis por verificar o documento e sinalizadores que indicam se o documento foi verificado, se está completo. Existe também um módulo, chamado conteúdo do documento (document content), onde são registradas as informações sobre histórico do paciente, sintomas, diagnóstico, tratamento entre outras.

A Figura 5 mostra o modelo de um objeto DICOM-SR e a especifica os Módulos do IOD Basic Text SR conforme a especificação do padrão DICOM A.35.3-1. Nela, encontram-se módulos do IOD (Information Object Definition) que possuem as informações e definições do objeto de SR, que define um módulo separado da árvore de conteúdo que contém esta informação (o SR Document General Module). Também pode-se observar, na 1Figura 5, que os SR - Relatórios estruturados são instâncias compostas DICOM assim como imagens e waveforms. Como tal, eles contêm atributos para identificar e descrever as entidades no modelo de informação composto. Isto inclui informações sobre o paciente, estudo, componentes do estudo, série e instância. Já que documentos SR são instâncias de classes SOP compostas, a eles são atribuídos identificadores únicos (Unique Identifiers - UIDs).

Instâncias compostas DICOM são normalmente persistentes, além do escopo de sua transmissão, elas são “documentos” ao invés de “mensagens”. Em um paradigma orientado a mensagens, elas podem ser transientes e serem descartadas após o uso. A informação de conteúdo e gerenciamento codificada em um relatório estruturado não pode ser alterada sem a criação de uma nova instância. Correções ou revisões precisam ser novas instâncias. A informação contida no módulo conteúdo do documento é dividida em "itens de conteúdo". Um "item de conteúdo" consiste de um par nome-valor, em que o nome é um código selecionado

de um dicionário de termos, e o valor é de um tipo dentre os quatorze tipos de valor definidos pelo padrão.

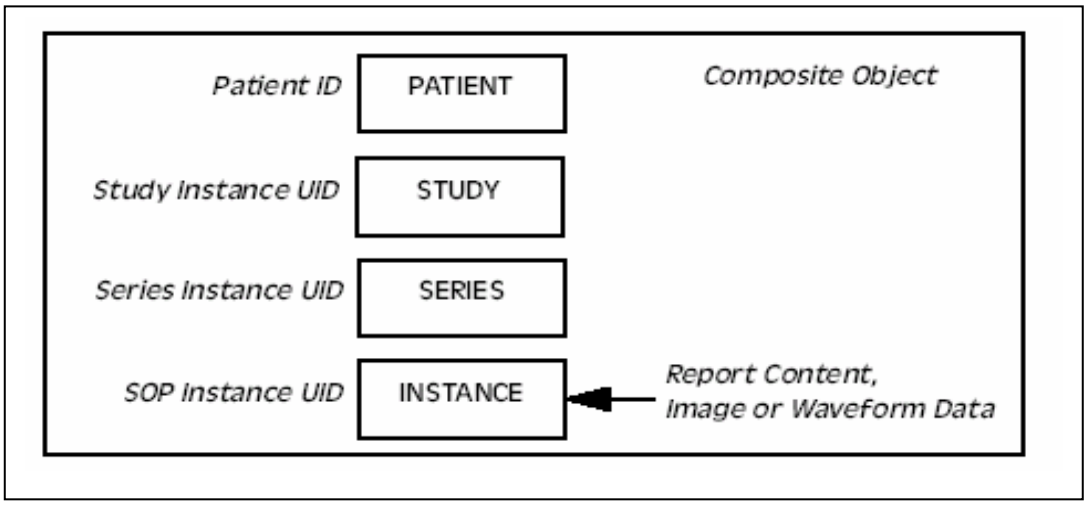


Figura 5 - Objetos compostos representam entidades múltiplas

Fonte: DICOM Strutured Report [CLUNIE2000]

Um dicionário de termos associa um nome de conceito humanamente significativo a um código. Dicionários amplamente utilizados são SNOMED para termos médicos, LOINC para observações clínicas e laboratoriais, e UCUM para unidades de medida. Entre os tipos de valor definidos pelo padrão para "itens de conteúdo" estão os tipos text (para texto), num (para números, porcentagem e outros), image (para imagens), date (para datas), e waveform (para formatos de onda, como eletrocardiogramas).

Todos os "itens de conteúdo" são organizados em uma hierarquia de informações, de modo que a informação nos níveis mais altos da hierarquia contém ou deriva de informações nos itens mais abaixo na hierarquia. Cada "item de conteúdo" (exceto o item raiz) contém um relacionamento, de um dos tipos definidos pelo padrão, com seu item pai de forma a evitar que o significado de um ramo da árvore seja ambíguo.

Quadro 4: Módulos da Classe Basic Text SRLN - A.35.3-1.

Módulos da Classe Basic Text SR - IE – Module – Usage –Meaning		
Módulos	Tipo	Conteúdo
Patient ID	Obrigatório	Atributos que identificam e descrevem o paciente que é o sujeito do estudo diagnóstico.
Specimen Identification	Obrigatório se é uma amostra.	Atributos que identificam uma amostra
General Study	Obrigatório	Informações que identificam e descrevem o estudo em que o documento sr está inserido, como por exemplo, identificador único do estudo, descrição do estudo, data e hora da realização.
Patient Study	Obrigatório	Informações sobre o paciente na data em que o estudo foi realizado, como, por exemplo, peso, e ocupação.
SR Document Series	Obrigatório	Atributos da série em que o documento está inserido, como por exemplo, identificador e modalidade da série (SR).
General Equipment	Obrigatório	Informações que identificam e descrevem o equipamento que produziu a série de imagens, como por exemplo, fabricante e nome da instituição a que pertence o equipamento.
SR Document General	Obrigatório	Informações que identificam o documento e fornecem o contexto em que o documento foi produzido, como por exemplo, nomes de pessoas responsáveis por verificar o documento e sinalizadores que indicam se o documento foi verificado, e se está completo.
SR Document Content	Obrigatório	As informações neste módulo estão organizadas em uma hierarquia e formam o conteúdo do laudo.
SOP Common	Obrigatório	Atributos que são necessários para o funcionamento e identificação da instância SOP associada. Não especificam nenhuma semântica sobre o objeto do mundo real representado pelo IOD, exemplo: Data em que a instância SOP foi criada.

Fonte: DICOM Strutured Report [CLUNIE2000]

O padrão DICOM SR especifica oito diferentes tipos de relacionamentos. Dentre eles estão contains (a informação do nodo pai está contida no nodo filho), has-properties (tem propriedades, a informação do nodo filho é uma propriedade da informação do nodo pai), has obs. Context (a informação no nodo filho é uma observação sobre a informação do nodo pai). As Figura 6 e mostram hierarquias exemplo.

Para cada classe SOP de Structured Report existem regras que determinam quais tipos de valor os itens podem assumir e quais tipos de relacionamento podem existir entre os diferentes tipos de itens.

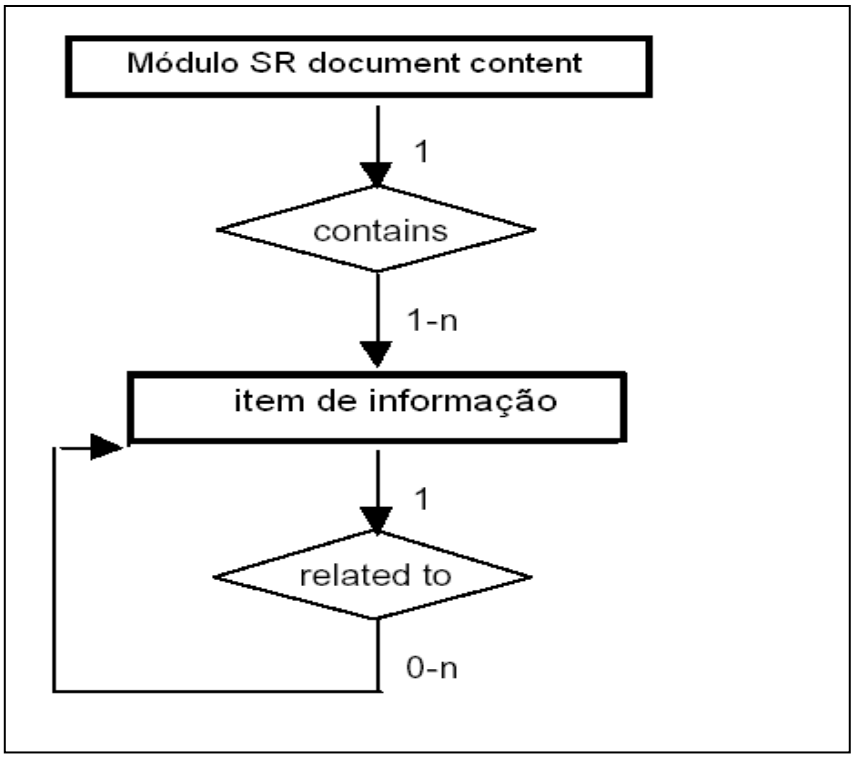


Figura 6 - Relacionamentos por referência

Fonte: DICOM Strutured Report [CLUNIE2000]

Em alguns casos, é desejável referenciar um "item de conteúdo" que já compõe outra parte da hierarquia, sem repeti-lo. Além do relacionamento com seu item pai, um item de conteúdo pode ter outro relacionamento dito "por referência" com outro item de conteúdo pertencente à hierarquia. Este tipo de relacionamento pode formar a hierarquia no formato de um grafo acíclico dirigido. Somente documentos da classe Comprehensive SR podem conter relacionamentos por referência. A 1Figura 7 mostra uma hierarquia exemplo da classe Comprehensive SR.

É importante poder especificar em um laudo, o contexto em que ele foi produzido. O contexto inclui quem ou o que produziu o conteúdo do documento, sobre quem ou a que um item se refere, e o procedimento ao qual aquela informação se refere. O contexto de observação é específico para relatórios estruturados e distintos de um contexto de aquisição (acquisition context), o qual se aplica a imagens e waveforms. O contexto de aquisição pode ser incluído em um

documento SR para codificar a informação presente num objeto DICOM tradicional “adquirido”, como uma imagem ou um waveform.

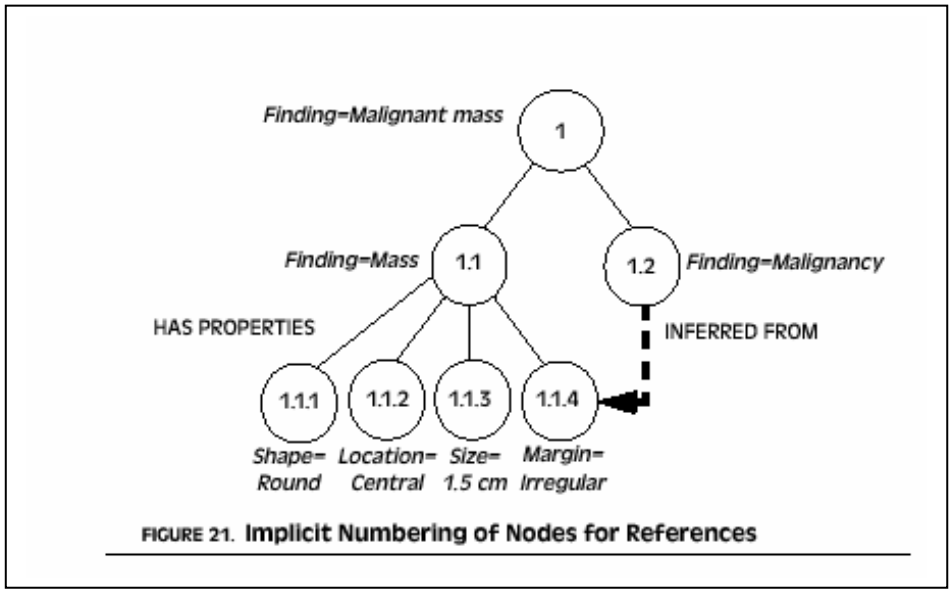


Figura 7 Hierarquia exemplo da classe Comprehensive SR

Fonte: DICOM Strutured Report [CLUNIE2000]

Para fazer este tipo de observação em um documento DICOM SR, adiciona-se um item contendo a observação ao item topo da hierarquia ou do ramo da hierarquia à qual a observação se refere. Desta forma, a informação no item de observação de contexto é válida para todos os itens desta hierarquia. O relacionamento entre o item topo e o item de observação deve ser do tipo has observation context. Diz-se que esta informação é “herdada” de itens descendentes deste um item, onde é adicionada a observação.

Pode ser necessário em um documento SR identificar o profissional que gerou o conteúdo ou parte do conteúdo de um laudo. Para este propósito, é feita uma observação do tipo observer context. Para especificar que determinado profissional gerou um braço da hierarquia, ao item topo desta hierarquia é adicionado um item filho cujo nome é “observador” e cujo valor seja o nome do profissional observador, com relacionamento do tipo “has observer context”. Um observador pode ser, um profissional que fez uma observação, um dispositivo ou

software que adicionou algum item de informação a um laudo, ou uma informação adicionada ao laudo por uma instituição de saúde.

O contexto de sujeito identifica e descreve sobre quem ou o que uma observação é feita. O sujeito da informação em um laudo é freqüentemente o paciente a que o laudo se refere. Mas parte de um documento SR pode ser, por exemplo, informações sobre alguma amostra retirada do paciente, ou um ou mais fetos encontrados em um exame.

O “contexto de procedimento” (“procedure context”) identifica e descreve “o que foi feito” Mais precisamente, fornece um meio de se especificar:

- “procedimento de aquisição de dados” deste documento;
- “procedimento de interpretação”.

O caso mais comum é um único relatório sobre um único procedimento de diagnóstico. Neste caso, o contexto de procedimento está associado às imagens, waveforms e a outros objetos compostos que são parte do mesmo estudo como identificado no General Study Module.

Se um relatório contém observações sobre mais do que um procedimento identificável ou resume informação sobre múltiplos procedimentos, então é necessário que ele defina sobre qual procedimento é uma observação.

Uma distinção é feita entre “procedimento de aquisição de dados” e “procedimento de interpretação”. Isto nem sempre é sinônimo. Em situações simples, os identificadores serão os mesmos. Isto ocorrerá quando o objeto SR object conter informações (tais como medidas) derivadas de imagens como parte do procedimento de aquisição ou quando o relatório é, na verdade, uma interpretação do “procedimento corrente”.

Na discussão de conteúdo e relacionamentos em seções anteriores, nenhuma referência de qualquer ordem particular foi feita sobre a ordem passagem pela árvore, nem de conceito de herança. Para a maioria do conteúdo, tais noções não são necessárias para definir o significado, além da definição avançada (“top-

down") de relacionamentos como "contains", "has properties" e "inferred from", as quais claramente estabelecem a fonte e o alvo do relacionamento. Mesmo no caso de cabeçalhos ("headings"), que são nomes de conceito dos itens de conteúdo de "container", a noção "containment" é suficiente para estabelecer uma hierarquia de cabeçalhos, subcabeçalhos contidos (internos), e assim por diante.

A exceção é o contexto de observação. Seria repetitivo ter que inserir cada item de conteúdo na árvore com seu próprio contexto de observação completo. Seria provavelmente inseguro deixar a propagação do contexto de observação indefinida ou implícita, desde que diferentes suposições de diferentes implementadores pudessem levar à ambigüidade. Portanto, são definidas regras para propagação explícita do contexto de observação, o qual:

- é herdado recursivamente por um nó, a partir de seu ancestral imediato;
- é propagado apenas por relacionamentos "by-value", e não por relacionamentos "by-reference".

No momento, o padrão declara que nunca se pode substituir um atributo de contexto de observação mais abaixo na árvore. Por exemplo, uma vez que o observador tenha sido estabelecido, nenhum dos descendentes do item de conteúdo (não importa a quão distante ele esteja) podem redefinir aquele atributo de contexto de observação novamente. No entanto, esta regra tem causado dificuldade e ninguém tem uma boa razão do motivo pelo qual ela foi introduzida. Por isso, é provável que esta restrição será removida e a substituição de atributos de contexto será permitida.

Já que relatórios estruturados DICOM são sempre objetos compostos ("composite objects"), já existe uma considerável quantia de contexto de observação presente no nível superior do conjunto de dados. Em particular, os módulos "Patient", "Specimen Identification", "General Study", "Patient Study" e "General Equipment" já contêm informação suficiente para identificar o assunto e o

procedimento. Em muitos casos, há informação para identificar o observador, particularmente quando o observador é um dispositivo e não uma pessoa.

Esta informação que vem “do lado de fora” da árvore de conteúdo (“content tree”) e referida como contexto “inicial” ou “default”. Ele não precisa ser repetido na árvore. O padrão declara que o contexto inicial de observação do lado de fora da árvore pode ser explicitamente substituído, ao invés de herdado, se for ambíguo.

Até este ponto, a hipótese tem sido que o contexto de observações é “direto”. Ou seja, o criador do documento SR (o observador) cria itens de conteúdo novos que descrevem o assunto diretamente. Isto não é sempre assim.

Algumas vezes, o criador do documento precisa citar um outro documento, tal como um relatório anterior, ou algo que foi contado a ele, tal como uma história falada pelo paciente. Além disso, um conteúdo citado “quoted content” pode também citar um outro material e assim por diante, recursivamente.

Não há uma forma de fazer as citações definidas no padrão. Isto é deixado aos “templates” e contexto de observação.

Os seguintes requisitos são comuns para documentos , citados e descrições faladas:

- o ponto na árvore no qual a citação começa, deve ser distinto;
- a pessoa ou dispositivo que faz a citação deve ser identificado;
- a pessoa ou dispositivo que são citados devem ser identificados.
- Em DICOM uma lista de códigos da qual um código pode ser escolhido é chamada “context group” ou grupo de contexto. Existem quatro possibilidades:
- não existem restrições, isto é, nenhum grupo de contexto é definido;
- uma lista de códigos é sugerida, mas códigos alternativos também podem ser usados. “baseline context group”.

- uma lista de códigos é especificada, mas a lista pode ser estendida com outros códigos, contanto que os novos códigos não tenham uma intercessão de significado com os códigos que já estavam na lista. “defined context group”;

- Uma lista específica de códigos deve ser usada e nenhum outro código deve ser usado, “enumerated context group”.

Os grupos de contexto não são definidos nos esquemas de codificação onde os códigos estão. Os grupos de contexto são definidos em um “Mapping resource” recurso de mapeamento que mapeia listas de códigos de onde são usados, no contexto para os esquemas de codificação em que estão definidos.

Laudos para domínio específico, como por exemplo, laudos de exame oftalmológico em determinada clínica, podem ter um formato comum.

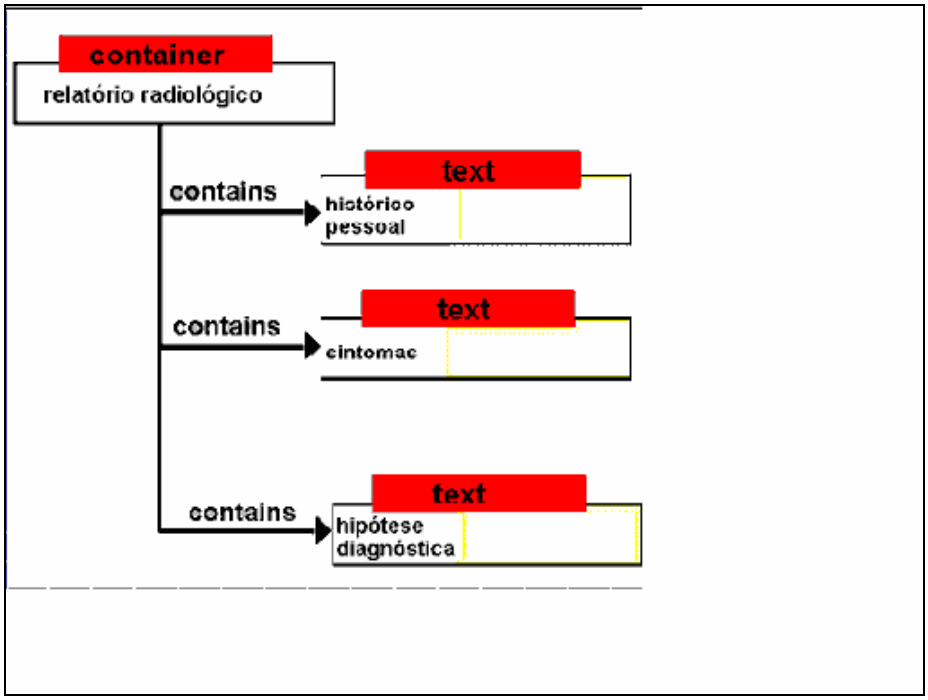


Figura 8 Exemplos de Relacionamentos em DICOM-SR

Fonte: [BORTOLUZZI2002]

O padrão DICOM SR permite que sejam usados modelos de laudos para aplicações específicas. Um SR template (modelo de SR) é um modelo de laudo

padrão que sugere ou restringe a hierarquia de itens de conteúdo ou parte desta hierarquia e que pode conter especificações de nomes (do par nome-valor), relacionamentos, tipos de valor e conjuntos de valores possíveis para um nome (do par nome-valor) [BELIAN2002].

Alguns exemplo disso podem ser vistos nas Figura 10, Figura 9, Figura 11 e Figura 11 referentes ao trabalho desenvolvido no Projeto Cyclops sobre um editor DICOM SR. [BORTOLUZZI2002]

No exemplo da Figura 9 está um a interface de montagem de modelos no. editor DICOM SR. [BORTOLUZZI2002]

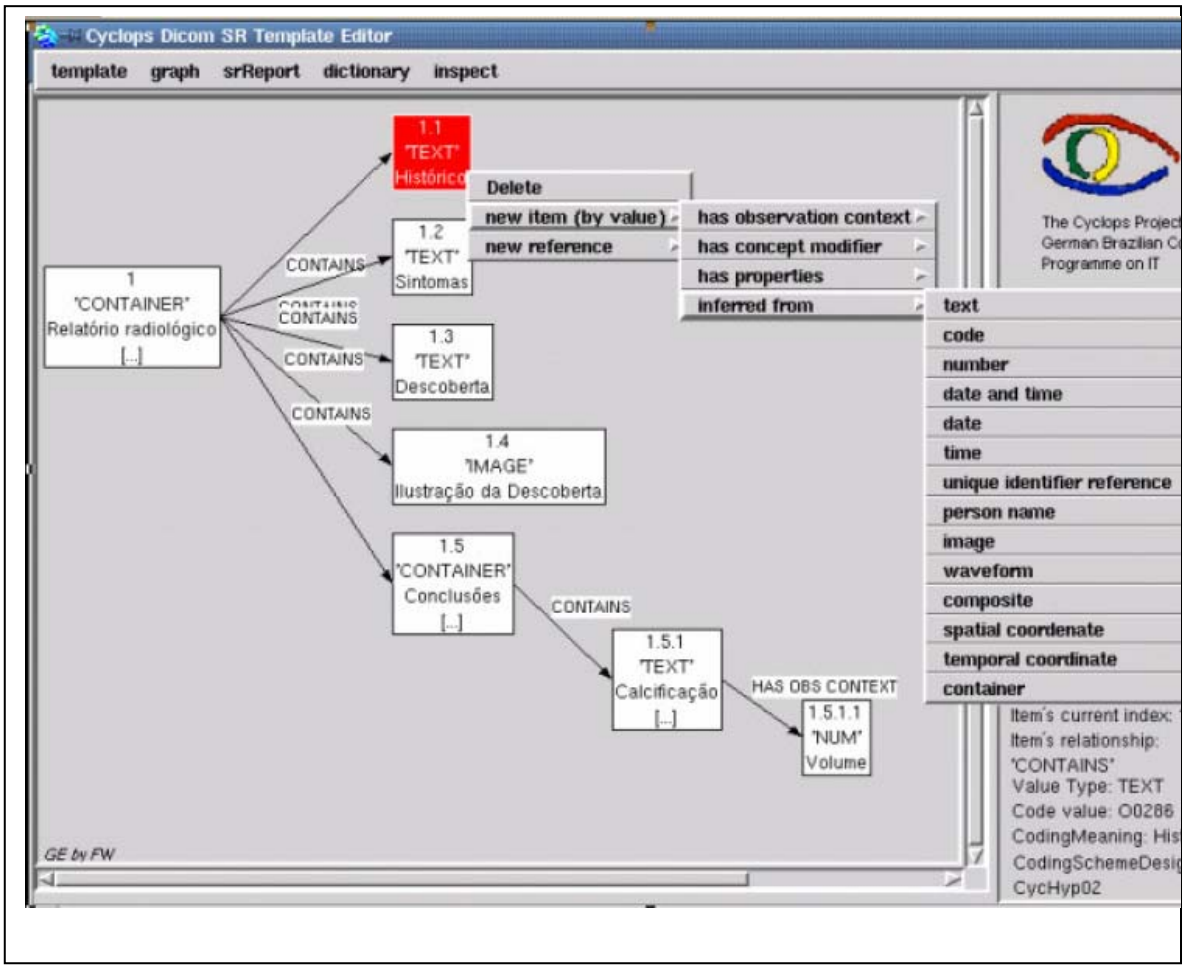


Figura 9– Interface de montagem de modelos no. editor DICOM SR.

Fonte: [BORTOLUZZI2002]

No exemplo da Figura 10. a estrutura de um Estrutura DICOM Strutured Report [CLUNIE2000], e na Figura 11 está um a interface de entrada de dados no. editor DICOM SR. [BORTOLUZZI2002]

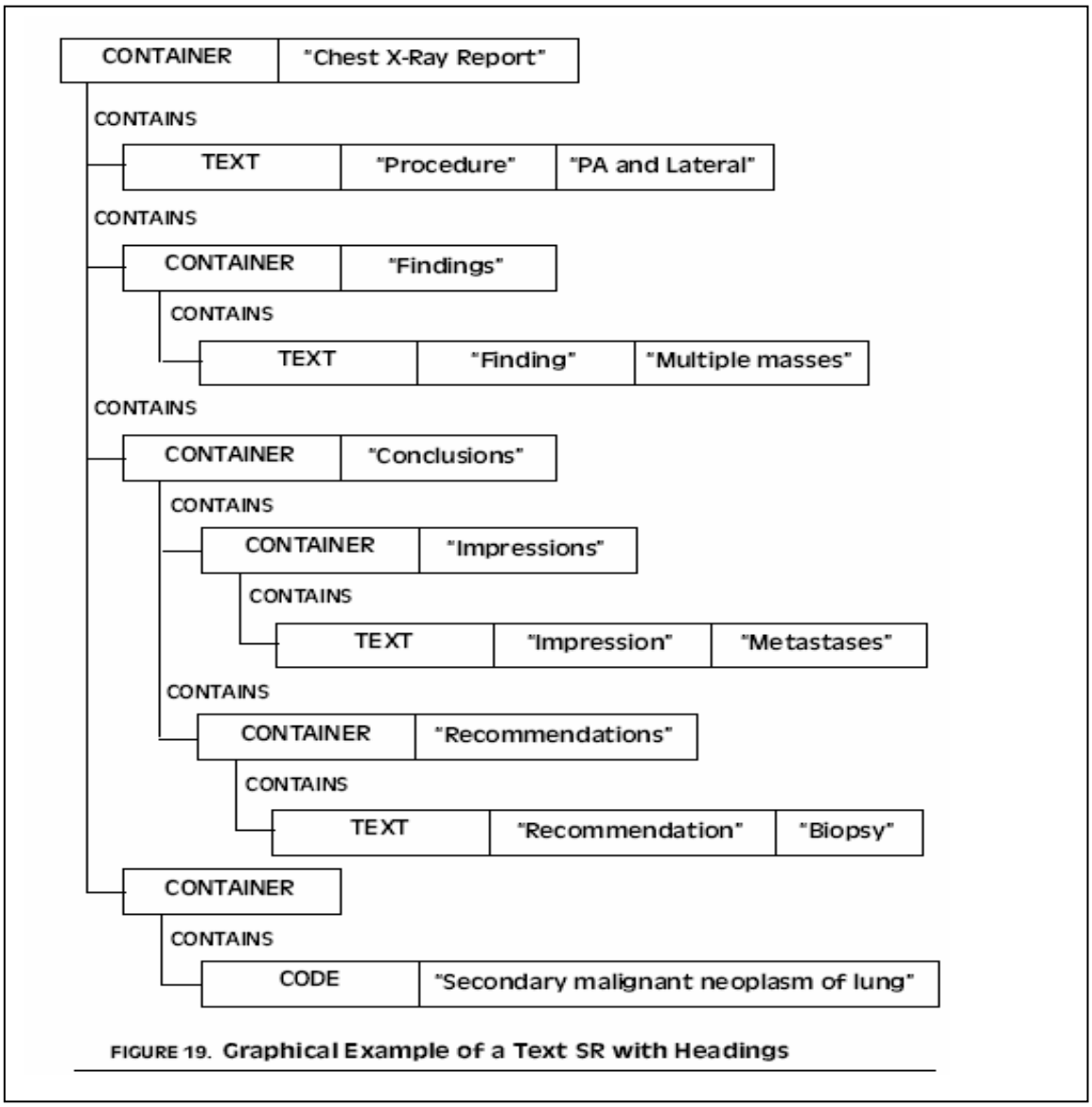


Figura 10– Exemplo de Estrutura DICOM SR

Fonte: DICOM Strutured Report [CLUNIE2000]

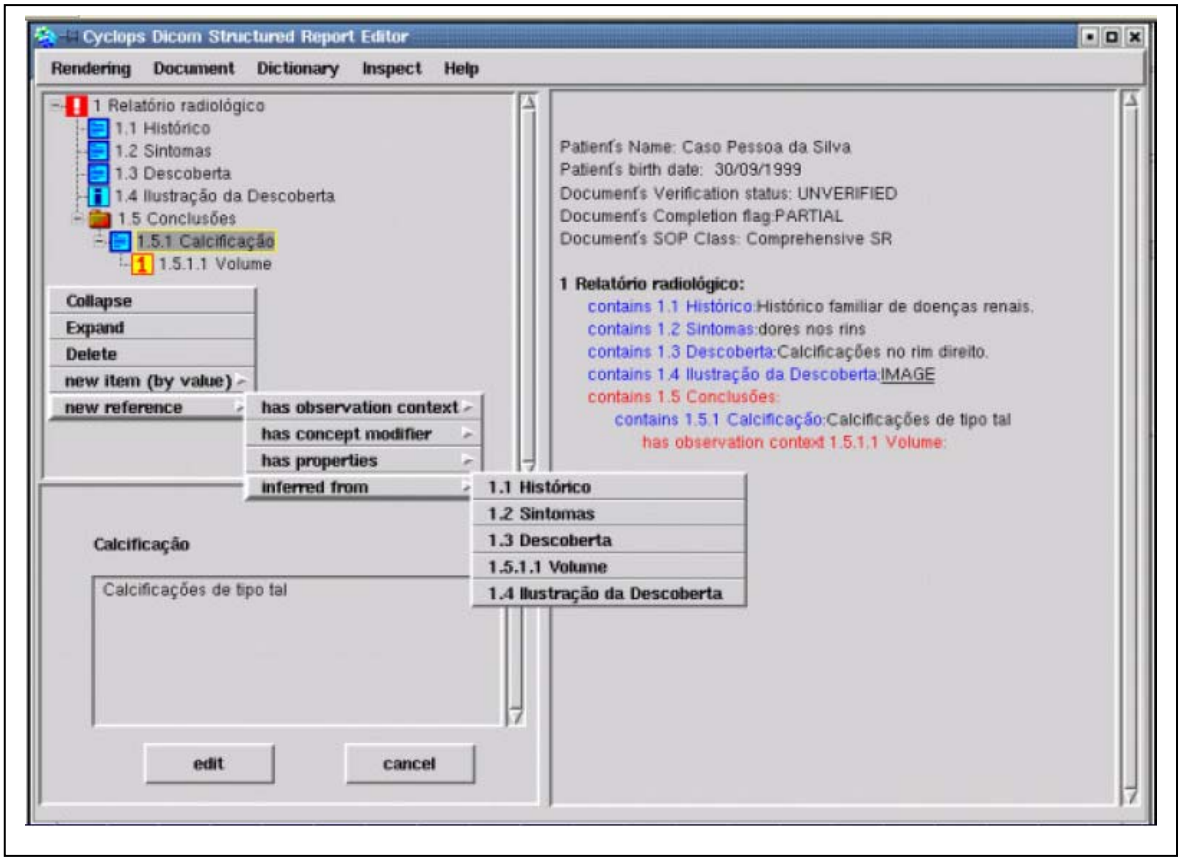


Figura 11 – Interface de entrada de dados no editor DICOM SR

Fonte: [BORTOLUZZI2002]

5 Análise detalhada de Requisitos

Neste capítulo, serão apresentados os requisitos detalhados para as funcionalidades e os diagramas de Caso de Uso.

5.1 Modelo do domínio do problema

Neste item, será apresentado o modelo às funcionalidades e os diagramas de Caso de Uso. Na Figura 12 pode-se ver a organização dos módulos do sistema aqui proposto em forma de organograma.

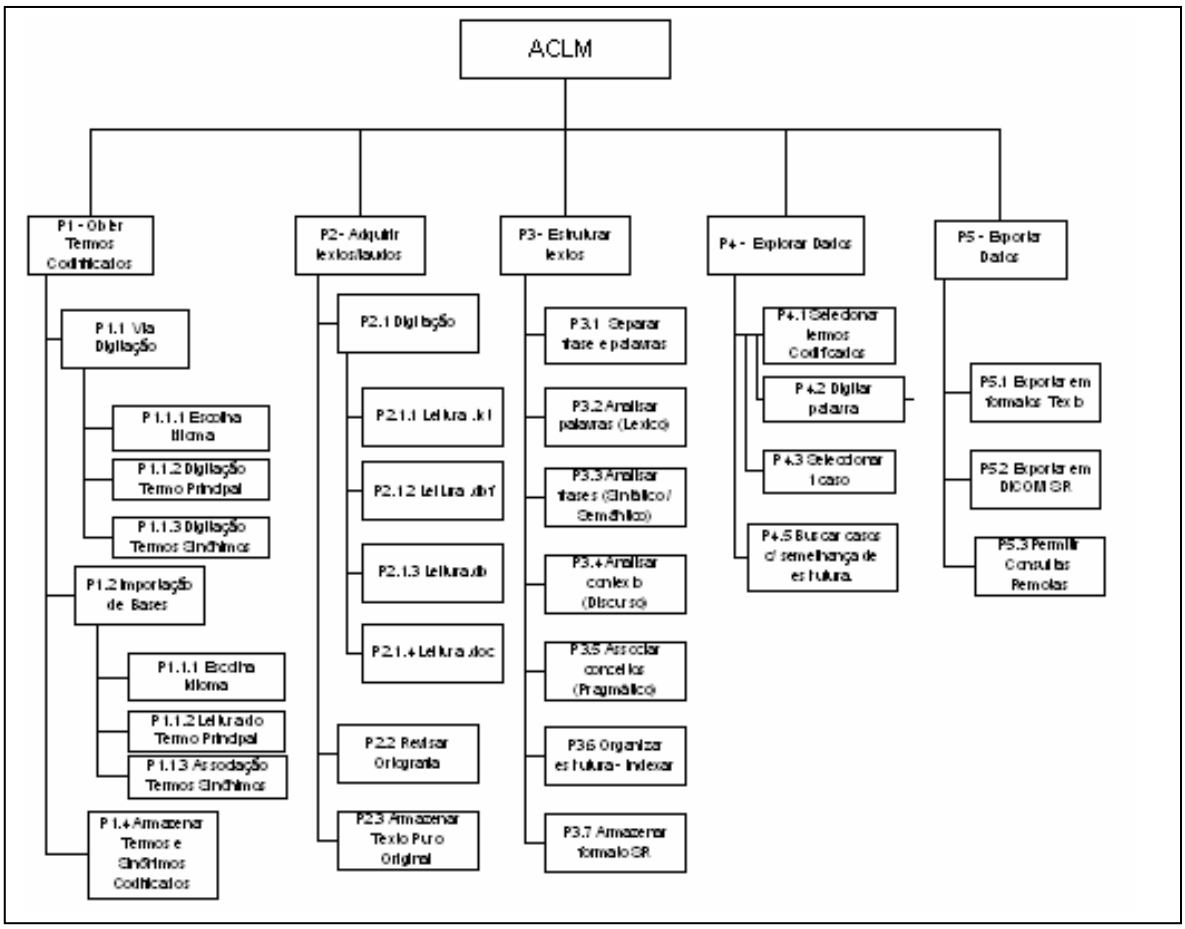


Figura 12 - Organograma

Fonte: Acervo do Autor

P 1. Obter Termos Codificados

Este módulo é responsável por controlar todo o gerenciamento dos termos médicos codificados, tais como, Snomed, CID-10 e outros. Para gerenciamento faz-se necessário um conjunto de controles do processo de aquisição e armazenamento de termos principais vinculados aos idiomas usados e também aos possíveis sinônimos e abreviaturas vinculadas a estes.

P 1.1. Via Digitação

O usuário deverá aqui, incluir os conjuntos de termos principais e sinônimos, bem como o respectivo idioma. Este submódulo é responsável pela digitação dos termos codificados. Este é um passo que se faz necessário para que os seguintes de estruturação e Exploração dos dados possam ocorrer.

P 1.1.1. Escolha do Idioma

O usuário deve poder escolher o Idioma em que deseja trabalhar, a indicação do idioma ao qual os termos serão incluídos é de fundamental importância para organização dos termos e permitir a posterior conversão entre diferentes idiomas.

P 1.1.2. Termo Principal

Permitir o cadastramento dos termos principais ou termos chaves de cada conceito da codificação utilizada. Esse termo principal é normalmente derivado diretamente do Latim, o que facilita o entendimento do mesmo, nos idiomas ocidentais por qualquer outro profissional da saúde, uma vez que é o idioma empregado para nomear descobertas científicas e também as partes do corpo, doenças, e outros termos sinônimos.

P 1.1.3. Termos Sinônimos

Cadastrar vinculado ao termo principal os termos sinônimos, que normalmente são termos de uso cotidiano dos profissionais da saúde de cada país e região, os quais, sem a devida vinculação, tornam-se de difícil manipulação e interpretação.

P 1.2. Via Importação de Bases

Este submódulo é responsável pela importação dos termos codificados de outras bases de dados, normalmente de padrão ASCII. Assim como a digitação, este é um passo que se faz necessário para que os passos seguintes de estruturação e exploração dos dados possam ocorrer. Os profissionais de Informática, em conjunto com o médico, deverão importar os conjuntos de termos principais e sinônimos bem como o respectivo idioma.

P 1.2.1. Escolha do Idioma

Indicar o Idioma em que o usuário deseja trabalhar sobre os Laudos, A indicação do idioma para o qual os Termos serão importados é de fundamental importância para organização dos termos e permitir o adequado armazenamento, assim como também posterior conversão de informação entre os outros idiomas.

P 1.2.2. Termo Principal

Importar os termos principais de outras bases de dados, exemplo do SNOMED, Topografia, Doença, Morfologia, Função, Agentes Químicos, Agentes Físicos, Organismos, Modificadores, Ocupação, Contexto Social e Procedimentos.

P 1.2.3. Termos Sinônimos

Importar os termos sinônimos vinculados aos termos principais.

P 1.3. Termos Auxiliares via Digitação

Permitir a digitação dos termos auxiliares que não são codificados. O Usuário médico deverá aqui, incluir os conjuntos de termos auxiliares e sinônimos bem como o respectivo idioma. Normalmente são qualificadores e quantificadores que serão associados aos termos médicos codificados principais e seus sinônimos..

P 1.4. Armazenamento dos Termos

Este submódulo é responsável pelo armazenamento dos termos principais e auxiliares, por idioma para permitir a posterior utilização do sistema de interpretação, bem como de orientação e consulta para outros sistemas.

P 2. Adquirir Textos de Laudos

Este módulo é responsável pela incorporação de novos laudos ou textos ao sistema, de forma a filtrar e prepará-los para a sua normalização de caracteres e palavras, e a correção ortográfica das mesmas. São antevistos dois tipos de normalização, porém outros ainda poderão ser abordados nas próximas interações do desenvolvimento deste documento.

P 2.1. Digitação

Este submódulo é pela digitação de um novo texto ou laudo para normalização e correção ortográfica, assemelhando-se a um editor de textos.

P 2.1.1. Leitura Arquivos padrão Texto (.txt)

Este submódulo é responsável pela importação de um tipo específico de arquivo o padrão (texto ou vulgo ".txt") que deve ser lido por qualquer tipo de computador, tendo características definidas pelos padrão ASCII de troca de arquivos. Deverá efetuar a normalização dos caracteres, eliminando tipos desconhecidos.

P 2.1.2. Leitura Arquivos padrão Data Base File (.dbf)

Este submódulo é responsável pela importação de um tipo específico de arquivo o padrão XBase (dbf), que irá requer desenvolvimento específico para cada nova origem de dados a sofrer a extração e recomposição de textos e informações em função da modelagem de dados, empregada para o desenvolvimento do sistema gerador das informações, normalmente faz-se necessária uma prévia conversão para arquivo texto.

P 2.1.3. Leitura de Tabelas em Bancos de Dados

Este submódulo é responsável pela importação de diversos tipos de banco de dados, podendo utilizar acesso direto ou via ODBC, que irá requer desenvolvimento específico para cada nova base de dados a sofrer a extração e recomposição de textos e informações em função da modelagem de dados, empregada para o desenvolvimento das tabelas do sistema gerador das

informações, normalmente faz-se necessária uma prévia conversão para arquivo texto.

P 2.1.4. Leitura Arquivos padrão Doc

Este submódulo é responsável pela importação de Laudos em um tipo específico de arquivo, o padrão doc, gerado pelo editor de textos Microsoft Word, sendo necessária uma conversão para arquivo texto.

P 2.2. Revisão Ortográfica

Este módulo é responsável pela revisão ortográfica dos textos, previamente adquiridos nas formas previstas no item P1, permitindo a interatividade do Médico com o sistema e sua avaliação dos termos que mais condizem com o significado do mesmo. Fará uso do conjunto de termos adquiridos e armazenados no item P1, ou em qualquer outro conjunto de termos definidos pelo usuário.

P 2.3. Armazenamento dos Textos

Este submódulo é responsável pelo armazenamento dos textos que já sofreram a normalização dos caracteres e correção ortográfica dos termos codificados (principais e auxiliares) por idioma, para permitir a posterior utilização do sistema de interpretação e estruturação prevista no módulo seguinte.

P 3. Estruturar Textos/Laudos

Este módulo é responsável por organizar as estruturas decorrentes da interpretação dos dados encontrados. Este módulo faz uso dos dados armazenados nos módulos P1 e P2.

P 3.1. Separar frases e palavras

Este módulo é responsável por selecionar o texto a ser processado e iniciar o desmembramento de palavras e frases nele contidas.

P 3.2. Analisar Palavras

Este módulo é responsável pela análise das palavras de forma Léxica, encontrando a possível relação com os termos codificados existentes.

Note que o sistema deve possuir, a priori, conjuntos de termos previamente definidos, obtidos a partir de uma população de interesse levantado com os médicos colaboradores deste projeto.

P 3.3. Analisar Frases

Este módulo relaciona-se com os aspectos Sintáticos e Semânticos das palavras dentro da frase, fazendo a identificação dos termos e suas funções dentro da frase, obedecendo às características de cada frase e procurando identificar e responder os itens do modelo de recuperação de informações proposto:

O que – achado de doença ou alteração;

Onde – região anatômica referenciada;

Como – qualificadores associados à região anatômica ou achados citados, Ex.: normais, alterados, lesionados, grande, pequena e outros.

Posição – qualificadores específicos vinculados à região anatômica que indicam a localização, que também sofrem combinações entre si. Ex. à direita, à esquerda, a frente, na superfície;

Note que este será um dos módulos de maior complexidade entre a saída do sistema e a opinião médica.

P 3.4. Analisar Contexto

Este módulo relaciona os aspectos Sintáticos e Semânticos das palavras entre todas as frases e subfrases, fazendo a identificação e relacionamento dos termos, que se referenciam a frases anteriores, normalmente permitindo completar as características não respondidas na mesma frase ou dentro do desdobramento das subfrases. Ex. Hilos e Mediastino Anatômicos;

Hilos -> Anatômico;

Mediastino -> Anatômico;

P 3.5. Analisar Conceitos

Este módulo relaciona-se com os aspectos Sintáticos e Semânticos e do Discurso de forma pragmática, permitindo a organização final dentro do texto a ser reestruturado, obedecendo às características de todas as declarações. Ex. Pulmões de formato Anatômicos. Leve formação de pneumonia à esquerda.

Pulmões -> Anatômicos

Pneumonia -> Leve -> Pulmão -> Esquerdo

P 3.6. Organizar a Estrutura

Este módulo relaciona-se com os anteriores do item P 3 e organiza a estrutura final que será armazenada no passo seguinte, criando a indexação necessária para facilitar a recuperação dos dados.

P 3.7. Armazenamento dos Textos

Este submódulo é responsável pelo armazenamento dos dados estruturados que passaram pelos processos previstos neste módulo, para permitir a posterior utilização do sistema de busca, recuperação prevista no módulo P 4.

P 4. Explorar os Dados

Este módulo é responsável pela exploração dos dados obtidos nos módulos anteriores. Tal exploração, o prevê o retorno de uma série de casos similares, assim como a uma porcentagem de similaridade do mesmo.

P 4.1. Selecionar Termos Codificados

Este módulo é a seleção direta de termos codificados com casos estruturados disponíveis, permitindo a sua rápida localização.

P 4.2. Digitar Palavras

Este submódulo relaciona-se diretamente com os aspectos relativos à interação do médico com o sistema, onde ele informará características desejadas dos casos que procura como as partes de um texto e as mesmas sofrerão os

processos descritos nos submódulos P 3.1 a P 3.6 e sendo, em seguida, Indexado para facilitar a busca dos casos semelhantes, de modo a tornar disponível para o sistema as informações médicas de que este procura.

P 4.3. Selecionar 1 Caso da Base

Este submódulo permitirá, através da escolha de um caso, que se refaça o procedimento de busca, levando em consideração todos os aspectos do caso selecionado, de modo a tornar disponível para o médico, outros casos similares ao por ele desejado.

P 4.4. Apresentar Resultados da Consulta

Este submódulo permitirá a visualização dos casos selecionados e possivelmente necessitará de maior detalhamento via processo de desenvolvimento incremental para possibilitar maiores facilidades ao médico a busca dos casos similares.

P 5. Exportar dados

Este módulo é responsável por gerenciar a exportação dos dados gerados pelo sistema, sejam eles arquivos de configuração de parâmetros, termos, textos ou laudos de forma estruturada ou original.

P 5.1. Exportar em formato Texto

Este módulo relaciona-se com a exportação dos arquivos no padrão texto ASCII (txt).

P 5.2. Exportar DICOM SR

Este módulo refere-se à exportação do formato DICOM Structured Report em formato compatível para posterior utilização por outros sistemas.

P 5.3. Permitir Consultas Remotas.

Este módulo cuida da geração de consultas que poderão ser enviadas de forma remota dos termos codificados.

5.2 Casos de uso

Neste capítulo apresentamos um caso de uso genérico do sistema proposto. Este caso de uso pode ser expandido em outros casos de uso mais específicos, etapa que ficará delegada para a segunda versão deste documento.

5.3 Caso de uso geral

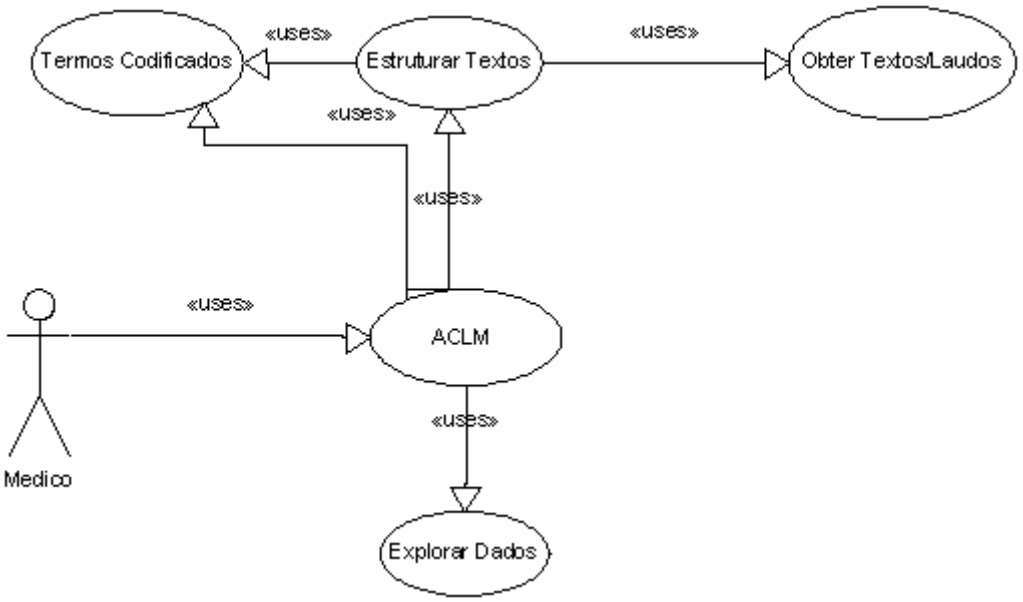


Figura 13 Caso de Uso geral

Fonte: Acervo do Autor

C.U. 1. Termos Codificados

1 Escolher Idioma do termo

Alternativas:

Verificar qual fonte dos termos gerados:

IF <digitado> VER : C.U. 1.1

IF <importado> VER: C.U. 1.2

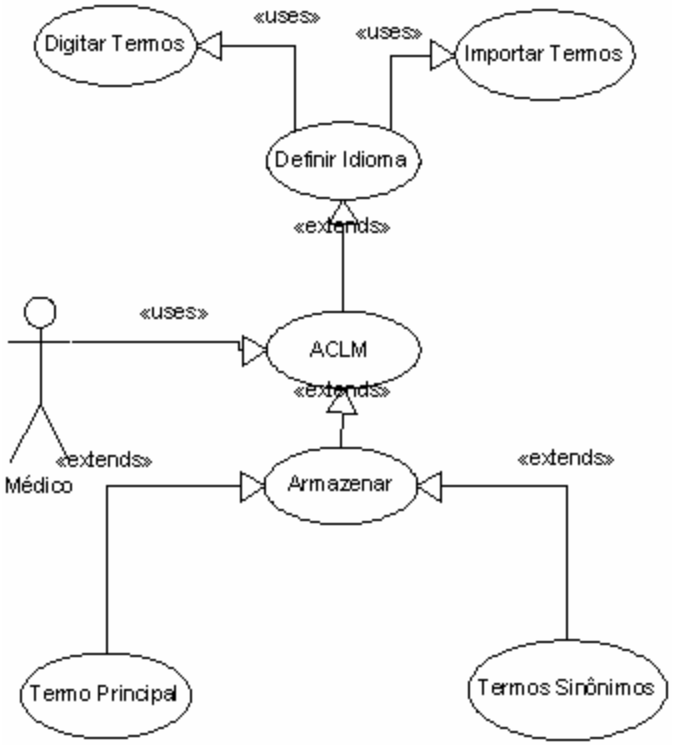


Figura 14 - Caso de Uso Termos Codificados

Fonte: Acervo do Autor

C.U. 1.1. Digitar Termos

Digitar Termo Principal

Digitar Referências Termo Principal

Digitar Termos Sinônimos

Digitar Referências Termos Sinônimos

C.U. 1.2. Importar Termos

Consistir arquivos

Importar Termo Principal, Referências Termo Principal

Importar Termos Sinônimos, Referências Termos Sinônimos

Listar resultado da Importação

Exceções:

O arquivo não existe – requisitar outro nome/caminho.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

C.U. 2. Obter Laudos

Digitar Laudo

Verificar para qual formato será importado

IF <TXT> VER: C.U.2.1

IF <DBF> VR: C.U. 2.2

IF <BANCO DADOS> VER: C.U.2.1

IF <DOC> VR: C.U. 2.2

Revisar Ortograficamente

Armazenar Texto Original na Base

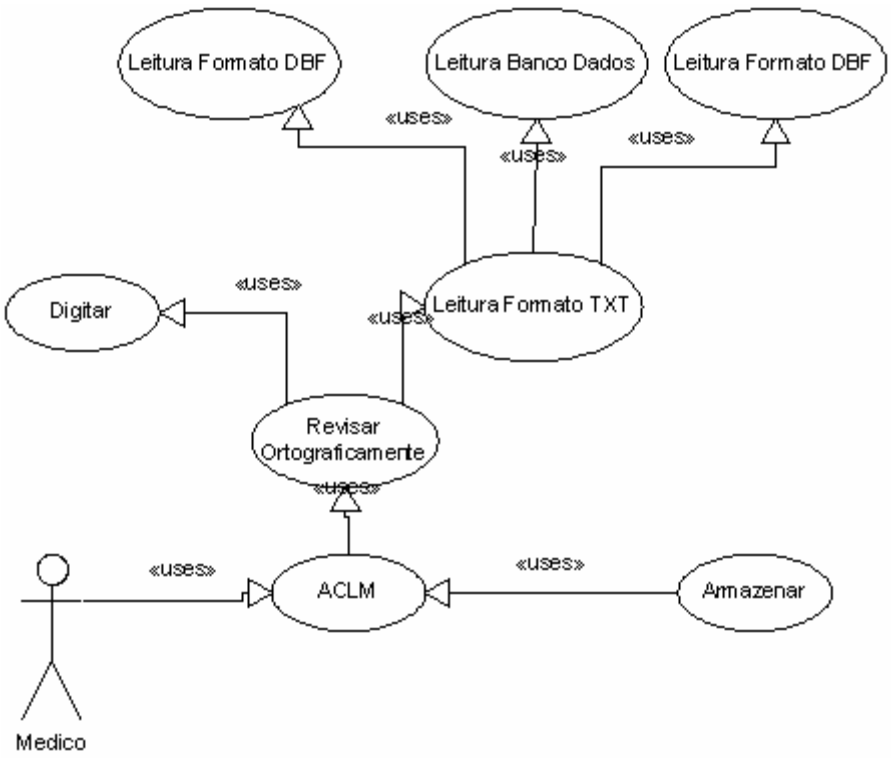


Figura 15 - Caso de Uso Obter Textos/Laudos

Fonte: Acervo do Autor

C.U. 2.1. Formato TXT

Consistir arquivos

Ler arquivo texto

Eliminar/Substituir Caracteres Especiais

Listar resultado da Importação

Exceções:

O arquivo não existe – requisitar outro nome/caminho.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

C.U. 2.2. Formato DBF

Consistir arquivos:

Ler Arquivo

Exportar dados em Padrão TXT

Efetuar Procedimentos VER C.U. 2.1

Listar resultado da Importação

Exceções:

O arquivo não existe – requisitar outro nome/caminho.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

C.U. 2.3. Formato Banco Dados

Criar Acesso via, ODBC

Informar Usuário e Senha

Selecionar Tabela

Ler Tabela

Exportar dados em Padrão TXT

Efetuar Procedimentos VER C.U. 2.1

Listar resultado da Importação

Exceções:

O Banco de Dados ou Conexão não existe – requisitar outro nome/caminho.

Usuário ou Senha inválidos requisitar outros.

Tabela não existe – requisitar outro nome/caminho.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

C.U. 2.4. Formato DOC

Consistir arquivos;

Ler Arquivo

Exportar dados em Padrão TXT

Efetuar Procedimentos VER C.U. 2.1

Listar resultado da Importação

Exceções:

O arquivo não existe – requisitar outro nome e caminho.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

C.U. 3. Estruturar Textos

Separar frases e Palavras VER P 3.1

Analisar palavras: Léxico VER P 3.2

Analisar Frases: Sintático - Semântico VER P 3.3

Analisar Contexto: Discurso VER C.U. 3.4

Associar Conceitos: Pragmático VER C.U. 3.5

Organizar Estrutura – Indexar VER C.U. 3.6

Armazenar no formato desejado VER C.U. 3.7

Exceções:

O caso já existente – requisitar outro caso.

Espaço insuficiente em disco – abortar a operação e informar o usuário.

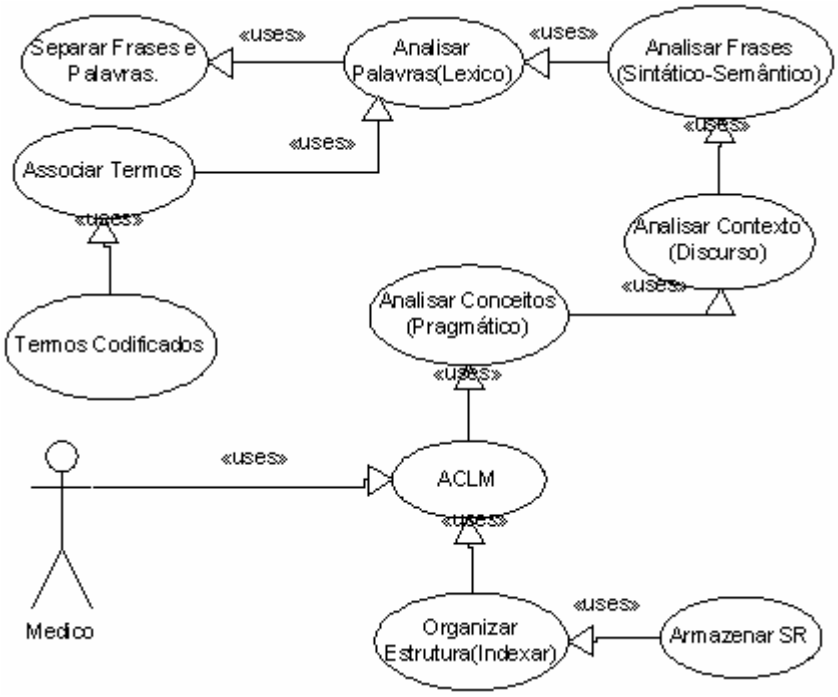


Figura 16 Caso de Uso Estruturar Textos

Fonte: Acervo do Autor

C.U. 3.1. Separar Palavras e Frases

Verificar Cadeias de String

Terminações:

Espaço em Branco – Fim da Palavra

ACL

Vírgula após Letra – Fim da Palavra

Pontos – Fim da Palavra e Frase

Ponto Final

Ponto e Virgula

Ponto de Exclamação

Ponto de Interrogação

Armazenar em Vetores, Coleções, etc...

C.U. 3.2. Léxico

Ler Palavras, Frases armazenadas.

Verificar Palavras nos Termos Codificados

Termo Encontrado

Alternativa:

IF <sim>, relacionar código Termo a Palavra.

IF <não>, buscar aproximação com outras palavras, informar o usuário.

C.U. 3.3. Sintático – Semântico

Verificar Estrutura da Frase

Organizar a Frase nos Itens:

O que (doenças, Morfologias, etc).

Onde (Anatomia)

Posição (Localização)

Como (adjetivos)

Verificar função das Palavras sem código

C.U. 3.4. Discurso

Dividir frase e termos concatenados nos Itens;

Formas novas frases com Complementos para:

O que (doenças, Morfologias, etc.)

Onde (Anatomia)

Posição (Localização)

Como (adjetivos)

Verificar função das Palavras sem código

C.U. 3.5. Pragmático

Verificar frase incompleta nos itens;

Buscar em outras frases Complementos para:

O que (doenças, Morfologias, etc.).

Onde (Anatomia)

Posição (Localização)

Como (adjetivos).

Verificar função das Palavras sem código

C.U. 3.6. Organizar Indexar

Ordenar Frases

Indexar

C.U. 3.7. Armazenar

Organizar Frases

Escolher Formato

Armazenar Caso

C.U. 4. Explorar Dados

Escolher Opção

Selecionar Termos

Digitar Palavras

Separar frases e Palavras VER P 3.1

Analisar palavras: Léxico VER P 3.2

Analisar Frases: Sintático - Semântico VER P 3.3

Analisar Contexto: Discurso VER C.U. 3.4

Associar Conceitos: Pragmático VER C.U. 3.5

Organizar Estrutura – Indexar VER C.U. 3.6

Selecionar um Caso Completo

Buscar Casos Semelhantes

Selecionar via Index

Apresentar Resultado

Exceções:

Nenhum casos de retorno – Informar Usuário

Muitos Casos – Pedir Usuário Número Máximo como limite

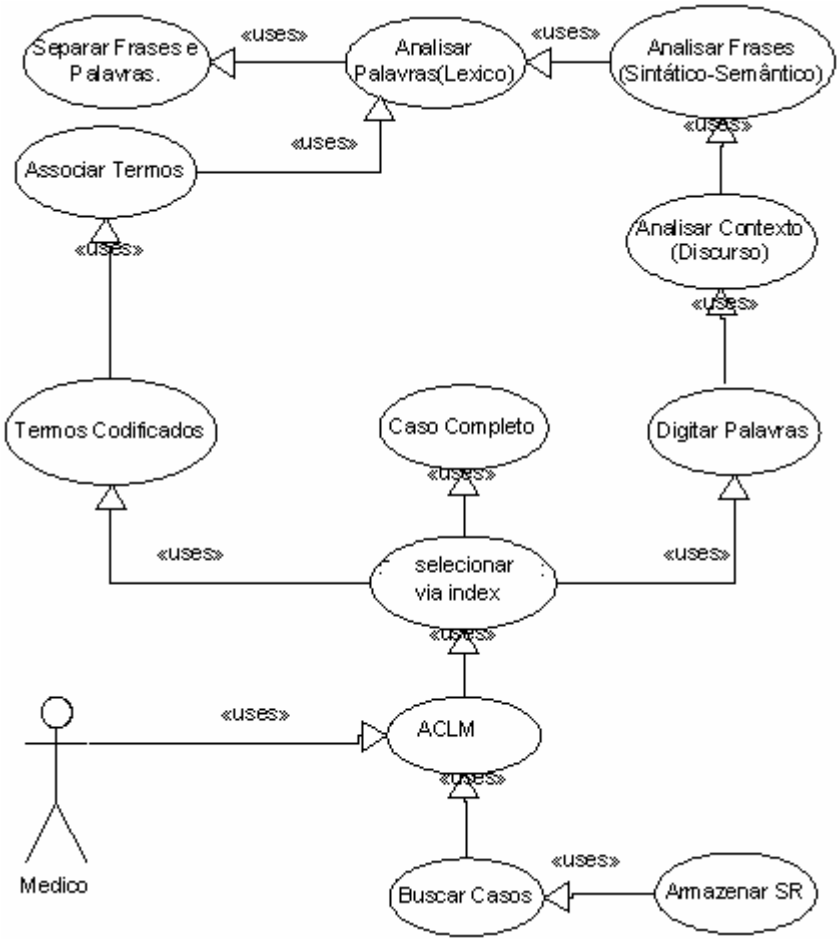


Figura 17 Caso de uso Explorar Dados

Fonte: Acervo do Autor

C.U. 5. Exportar Dados

Escolher Formato Padrão

Formato TXT

DICOM SR

Consultas Remotas aos Termos

Construir Conexão

Usuário e Senha

Realizar Consulta

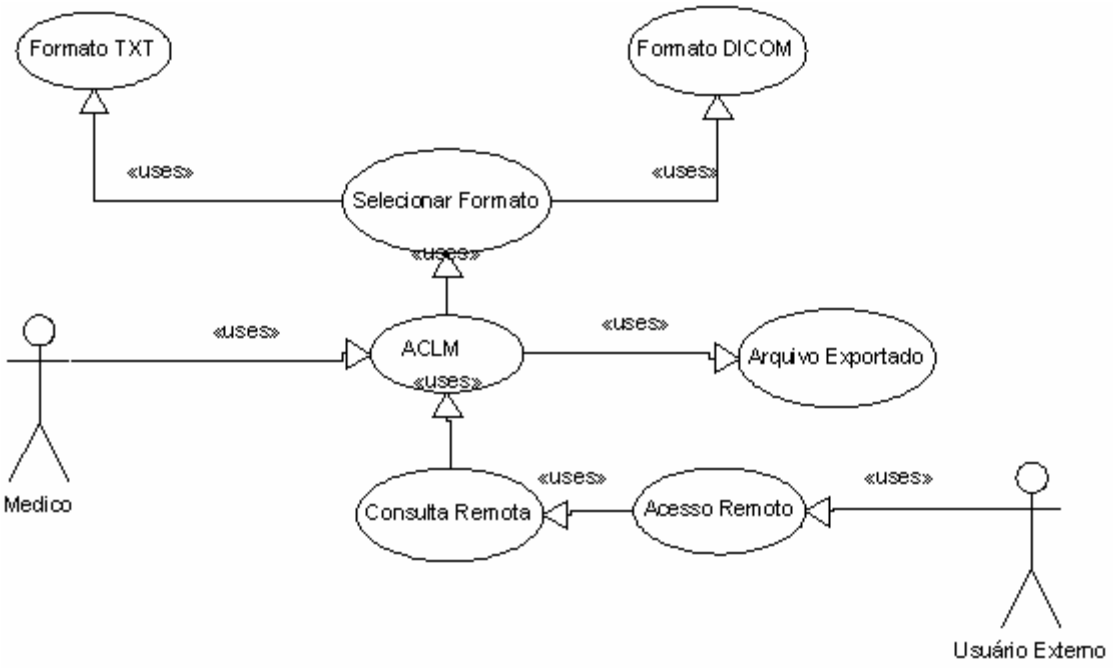


Figura 18 - Caso de uso Exportar Dados

Fonte: Acervo do Autor

5.4 Requisitos funcionais

R.F. 1. ***O sistema deve permitir a aquisição e controle de Termos Principais, Auxiliares e seus Sinônimos Codificados em Padrões Internacionais.***

R.F. 2. ***O sistema deve permitir a aquisição de novos casos.***

R.F. 3. ***O sistema deve permitir consultas por termo Codificado, combinação de palavras (texto livre), escolha de Caso.***

5.5 Requisitos não funcionais

R.N.F. 1 ***O sistema deve definir um formato de arquivo para exportar os dados.***

R.N.F. 2 ***O Sistema deve possuir uma Interfaces para integrar-se aos softwares do Projeto Cyclops..***

6 Modelo de Análise de Laudos

Como a linguagem médica empregada nos Laudos Radiológicos é sucinta e possui algumas características particulares encontradas, tanto textos dos laudos em Alemão, como em Português, como empregar apenas frases afirmativas ou negativas, normalmente sem o emprego de verbo.

Este recurso empregado nos textos dos laudos, foge às características lingüísticas de textos literários ou jornalísticos em ambos os idiomas, porém trazem benefícios da eficiência e rapidez na leitura e interpretação por parte dos especialistas a quem se destinam os exames realizados.

Os laudos gerados nos exames têm também características e vocabulários próprios, seguindo uma estruturação específica por tipo de exame. Habitualmente são destacadas apenas as áreas anatômicas de interesse do estudo. Por exemplo, em uma tomografia da região do tórax são destacadas regiões anatômicas como mediastino, Traquéia, Brônquios, Pulmões Hilos e Pleuras.

Para efetuar a análise do laudo radiológico, baseou-se em informações extraídas no livro de radiologia alemão CT-und MRT- Normalebefunde [MÖLLER1998] e no livro SNOMED Systematisierte Nomenklatur der Medizin [WINGERT1984]. A SNOMED contém desde as regiões anatômicas, ocupação, agentes químicos, físcos, vírus, além de possíveis doenças, tal como inflamações, deformações e outros achados significativos para os especialistas médicos.

Durante as análises identificou-se a estrutura básica dos laudos, que normalmente, possui partes distintas iniciando a identificação do paciente e do médico requisitante além de constar evidentemente o tipo de exame realizado com a respectiva região anatômica ou topológica do corpo onde foi realizado. A seguir, a técnica do Exame empregada citando quando o caso contrastes e seqüência de imagens, os dados clínicos contêm doenças conhecidas e reclamações e queixas do paciente com poucas frases e algumas patologias previamente conhecidas de exames anteriores, como por exemplo, “Suspeita de Pneumonia”.

No item seguinte “achados” ou “corpo do laudo”, fica descrito o que estaá sendo visualizado no exame realizado. Utilizando-se apenas dados importantes ou alterações sobre a região anatômica, assim como se utiliza frases curtas com o baixo emprego de verbos, que seguem normalmente característica de afirmar ou negar alterações, presença de corpos estranhos ou doenças. O parecer ou conclusão descreve o que existe de mais importante visto no exame realizado, normalmente repete algumas das frases contidas nos achados.

O Rodapé identifica o médico radiologista responsável pelo laudo como pronome tratamento: Dr. (Doutor) e nome completo e número de registro.

6.1 Modelo de Recuperação e Interpretação de Laudos

Após analisar diversos laudos, tanto no idioma Português quanto no idioma Alemão, juntamente com médicos, foram montados diversos modelos de recuperação. Inicialmente fez-se uso de regras específicas para cada tipo de exame, o que conduziu a alguns fracassos, devido à complexidade e dificuldade de expressa-los na forma um grande número regras computacionalmente.

Então, se passou a estudar as semelhanças e diferenças entre os diversos laudos e a estrutura das frases contidas nesses. Empregou-se a análise da freqüência das palavras, apesar da diversidade de frases tanto em número dessas assim como de palavras, além das formas das construções e posicionamento dentro da frase das mesmas. Verificando a ocorrência das mesmas observou-se a existência de um padrão que poderia ser repetido no todo e em partes referindo-se a função das palavras dentro das frases em diferentes tipos de exames.

A partir dessa análise montou-se um modelo, que demonstrou ser bastante eficiente, baseando nas informações mais importantes a encontradas, que foram as regiões Anatomia (onde?) as Doenças ou alterações identificáveis (o que?) e a qualificação de cada uma delas (como? e Como (do o que)?). Isto se pode verificar na Quadro que traz as informações chaves a serem extraídas de uma das linhas dos laudos que estão sendo analisados.

Quadro 5: Informações chave a serem extraídas

INFORMAÇÕES CHAVE A SEREM EXTRAÍDAS		
Informação	Conteúdo	Exemplo de Valores
Onde.	Anatomias	Pulmões, Hilus, Traquéia....
Posição.	Área dentro da anatomia	Direita, esquerda, lateral, superfície
Como	Situação da Antomia	Normal, sem alterações, anatômicos
O que.	Doenças ou alterações identificáveis.	Pneumonia, Sinusite ....
Como (do oque)	Situação da doenças ou alterações identificáveis.	Crônica, Aguda, Leve, ausente ....

Fonte: Acervo do Autor

A região anatômica (Onde) pode ser interpretada em SNOMED sob o eixo Topografia “T-Topographie“. Outra figura importante agregada à região anatômica, surgiu nos processos de implementação do modelo, a “posição”, que também pode ser convertida em SNOMED como subclasse do eixo geral “G”, especificamente o “G-A“ que são ligações como: direita (G-A100), esquerda (G-A101), bilateral (G-A102) e outros. Este tipo de termo também pode estar combinado com proposição de lugar tais como: em, na / no, acima, abaixo. Ambas podem ser identificadas pelos itens ou perguntas onde e qual posição. Ex. Onde = Pulmão e Qual posição = direita

Encontram-se também palavras que especificam problemas encontrados, essas podem ser identificados pelo item ou pergunta “o que”, e também SER convertida em SNOMED em diversos tipos tais como, Morfologias “M-Morphologie“, Doenças „D-Krankenheit“ ou Disfunções „F-Funktion/Disfunktion“.

Associados às regiões anatômicas e / ou aos problemas “Qual”, encontram-se os termos qualificadores identificados pelo item ou pergunta “Como”, que pode conter diversas formas, tais como normal, intacto, anatômico, etc., uso de negação associado como sem, ausente e outros, que indicam a ausência de problemas. Ou outros como mínimo, leve, pouco, que indicam pequenos problemas ou ainda forte, agudo, crônico e outros, que indicam grau mais avançado de um problema.

Baseando-se na análise detalhada de requisitos, realizada com os médicos, iniciou-se o desenvolvimento com a leitura dos laudos radiológicos em formatos

como documentos (.doc) e padrão XBase tranformando-os em padrão texto (.txt) eliminando cabeçalhos de padrões de formatação.

Na correção ortográfica, usou-se a ferramenta GNU Aspell[4], um software livre que se encontra disponível em vários idiomas e sistemas operacionais e permite a criação de novos dicionários de dados. No processo de recuperação das informações na análise sintática e semântica, foi necessária a criação de uma classificação das palavras pela sua classe gramatical Substantivo, Artigo, Pronome, Advérbio, Verbo, Numeral, Adjetivo, Conjunção, cuja finalidade é auxiliar na estruturação das informações encontradas.

Após os testes seguintes do modelo de classificação, encontrou-se uma quantidade expressiva de adjetivos que eram derivados de substantivos e que por isso possuíam grande importância para a organização e interpretação das frases, necessitando, assim, de um tratamento especial com os mesmos. Eles fazem referência direta a anatomias codificadas em SNOMED, como por exemplo, “superfície pulmonar” que pode ser interpretada como superfície do pulmão código T-28000, como por exemplo na Figura 19 modelo de Tela da de Cadastro de SNOMED.

Figura 19– Exemplo de Cadastro Termo Snomed

Fonte: Acervo do Autor

Um dos maiores problemas nesta fase foi a falta de SNOMED no idioma Português, porém com a ajuda dos preceptores e residentes da Residência Médica do Hospital regional Alto Vale foi possível traduzir os principais termos utilizados nos

testes realizados, baseando-se no livro SNOMED Systematisierte Nomenklatur der Medizin em alemão[WINGERT1984], onde normalmente a primeira referência de cada termo está no latim, em função da grande influência do mesmo na língua alemã assim como na designação de termos científicos

A proposta do processo de análise dos Laudos Textuais prevê as seguintes etapas, representando sob a forma de diagrama na Figura 20, iniciando com a Leitura e verificação dos Laudos e a Recuperação dos textos que podem estar em diferentes formatos (Word, Dbf, Txt e banco de dados) transformando-os para padrão texto ASCII. Procede-se também a limpeza dos textos dos caracteres que não pertencerem ao conjunto válido de caracteres vinculados aos textos radiológicos.

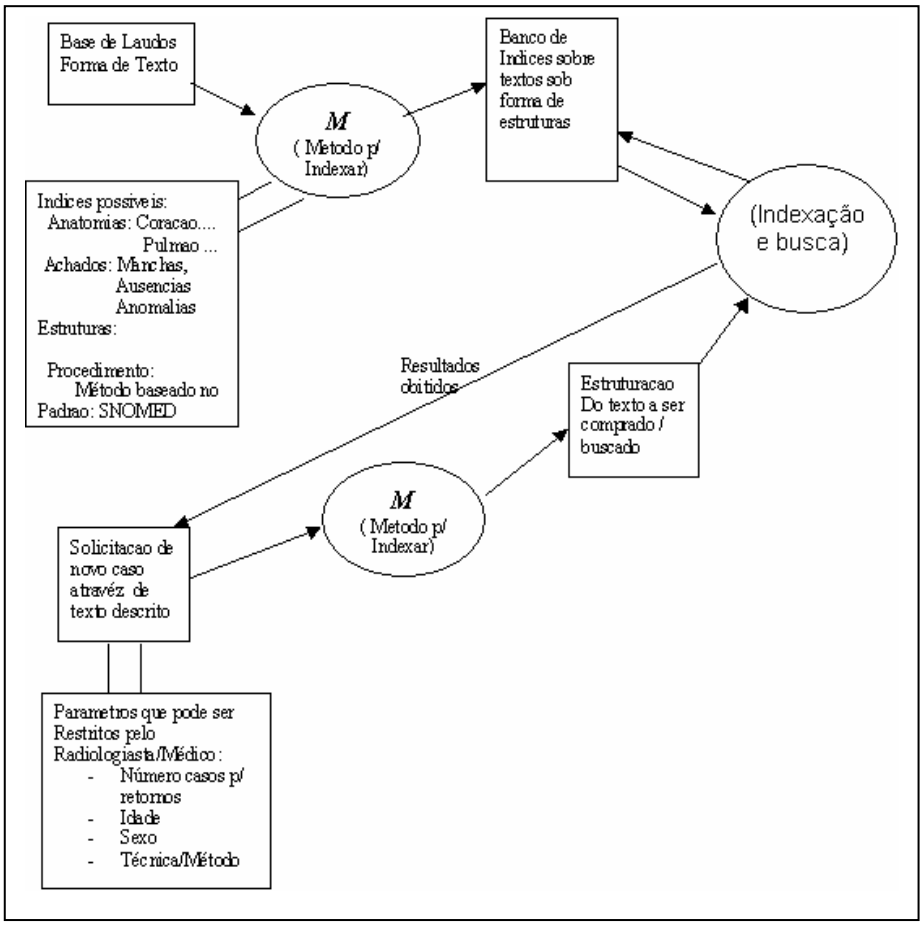


Figura 20 Diagrama Fases de Recuperação e Interpretação de Laudos Médicos

Fonte: Acervo do Autor

Para verificação ortográfica do texto, utilizou-se a ferramenta própria para este fim, a seguir a utilização dos métodos propostos (representado pela letra M), armazenamento da informação de forma estruturada e indexação e busca de casos que contenham estruturas. Na fase seguinte, foi realizada a geração das estruturas DICOM SR com base nas informações estruturadas armazenadas.

A forma de obtenção dos Laudos a serem utilizados que provêm do resultado do processo entre médicos especialistas nas mais diversas áreas dentro da medicina, mas neste caso específico, apenas das clínicas Radiológicas foi descrito no capitulo 2.

Após o processo de limpeza dos laudos recebidos onde são retirados os “ruídos” ou “sujeiras” como caracteres inválidos e /ou não pertencentes aos conjuntos de letras, números, caracteres especiais de acentuação, pontuação, operadores matemáticos e outros utilizados no dia-a-dia.

No processo seguinte, faz-se necessária a verificação ortográfica das palavras minimizando assim, problemas no processo de interpretação. O passo seguinte é a divisão dos Laudos em frases numeradas pela ordem de leitura e subseqüentemente a divisão destas frases em palavras e a substituição das palavras que constarem na lista de palavras equivalentes. Que tem por finalidade substituir determinadas palavras que os especialistas médicos julguem inadequadas ou de significado semelhante por uma palavra central melhorando o texto. Principalmente no idioma alemão onde a construção de uma frase permite utilizar o genitivo expressando dois termos separados utilizando o equivalente ao português da preposição de mais os artigos ou simplesmente juntar as duas palavras formando uma única.

Para o próximo passo: a análise léxica e sintática é necessária à construção de um dicionário destas palavras classificando as palavras e em todas as suas inflexões, (Como exemplo desta classificação no Quadro 6), tal como: artigos, substantivos, adjetivos, verbos, numerais, outros.

Quadro 6: Exemplo Classificação Palavras no Dicionário.

Palavra	Tipo	Número	Gênero	Situação	Relativo a Anatomia	SNOMED
HILO	Substantivo	Singular	Masculino			T28080
HILOS	Substantivo	Plural	Masculino			T28080
Pulmonar	Adjetivo	Singular	Ambos		T28000	
Pulmonares	Adjetivo	Plural	Ambos		T28000	
Anatômico	Adjetivo	Singular	Masculino	Normal		
Anatômicos	Adjetivo	Plural	Masculino	Normal		
Anatômica	Adjetivo	Singular	Feminino	Normal		
Anatômicas	Adjetivo	Plural	Feminino	Normal		

Fonte: Acervo do Autor

Neste momento, as palavras passam a ser então comparadas com formato e avaliadas conforme a classificação a elas atribuídas de onde então, começa-se a verificar a sua função dentro da frase, ou seja, análise sintática utilizando algumas regras como exemplo entre substantivos e adjetivos na frase abaixo:

Ex. Frase:

HILOS PULMONARES ANATÔMICOS

Na Posição 1, está o substantivo Hilos com o código SNOMED T28080, na posição 2, está o adjetivo Pulmonares, que por ser relativo ao substantivo pulmão tem o código SNOMED T28080, e na posição número 3, o adjetivo anatômicos que possui a situação normal atribuída pelo significado de estar em conformidade com a anatomia do local a que se refere.

No item seguinte, fazem-se necessárias as análises sintática e semântica que permitem o uso de regras como as descritas abaixo:

Se o Adjetivo de posição N = Plural e Substantivo Posição N – 1 Plural refere-se diretamente ele.

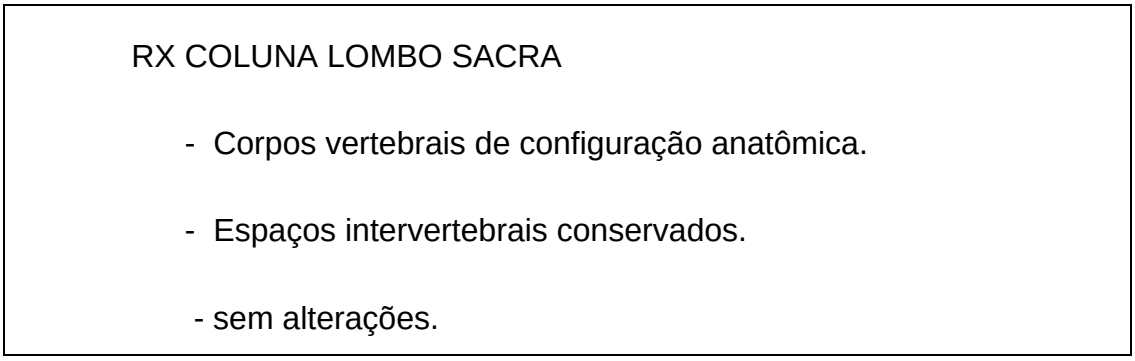
Se o Adjetivo de posição N = Plural e Substantivo Posição N – 1 Singular refere-se a ele e a outros Substantivos Posições N -2, N–3.

Se o Adjetivo de posição N = Plural e existir somente 1 Substantivo Posição N – 1 Singular = erro de escrita, mas refere-se diretamente ao Substantivo Posições N –1.

Se o Adjetivo de posição N = Singular e Substantivo Posição N – 1 Singular = refere-se diretamente a ele.

Se o Adjetivo de posição N = Singular e Substantivo Posição N – 1 Plural = erro de escrita , mas refere-se diretamente ao substantivo posição N –1.

Surgiram também situações diferentes, onde se fez necessária a análise de discurso como, por exemplo a Figura 21, onde o laudo tem frases que não possuíam nenhum termo anatômico e foi necessário recorrer ao termo indicado na primeira linha do laudo, onde se encontra o procedimento RX e o termo principal Coluna Lombo Sacra..


RX COLUNA LOMBO SACRA

- Corpos vertebrais de configuração anatômica.
- Espaços intervertebrais conservados.
- sem alterações.

Figura 21 - Rx Coluna Lombo Sacra

Fonte: Acervo do Autor

6.2 Exemplos do Modelo de Recuperação e Interpretação de Laudos

Um exemplo da análise dos dados pode ser visualizado no Quadro 7, abaixo, onde, além da análise realizada, estão demonstradas situações de geração de frases estruturadas visando a gerar, posteriormente, laudos no padrão DICOM SR.

Quadro 7: Exemplo de Análise Laudo

Exemplo de Análise de Laudo				
Frase Original	Onde		Como	Número Frases Geradas
	Anatomia (Substantivos)	Snomed	Qualificação (Adjetivos)	
Traquéia, brônquios principais e lobares, de calibre normal.	Traquéia	T25000	Calibre Normal	3
	Brônquios Principais	T26000		
	Brônquios Principais	T26200		
Hilos pulmonares anatômicos	Hilos Pulmonares	T28080	Anatômicos	1
Superfícies pleurais anatômicas e sem alterações	Pleura Superfície = Posição	T29000 GA168	Anatômicas Normal (Sem alterações)	2

Fonte: Acervo do Autor

Na primeira frase original do quadro, existem três regiões anatômicas e uma qualificação o que geraria três frases conforme os itens Onde e Como, em uma relação N anatomias para 1 qualificador, além da palavra “brônquios” servir também a lobares pérvios. Na segunda frase, nota-se a relação de 1 para 1 e na última, além de uma relação anatomia 1 para N qualificadores, além de encontrar o adjetivo derivado pleural que necessita ser traduzido para Pleura e superfície identificar a posição na pleura cujo código é G-A168.

Foram comparados aproximadamente 40 Mb em textos de laudos do Hospital Regional Alto Vale e a Clínica DMI e tratados essencialmente os laudos de tórax que primeiramente sofreram a correção ortográfica e foram desenvolvidas diversas rotinas para as analises das frases, palavras e a formação de estruturas

iniciais de informações. Foram realizadas também algumas análises de laudos em Alemão da Radiologische Gemeinshaftspraxis Buddenbrock, Blasinger und Benz conforme a Quadro 8. Estes resultados preliminares estão em análise por médicos, as informações geradas pelo protótipo com objetivo de integrá-las, no passo seguinte, aos softwares do padrão DICOM SR que se destaca por fazer parte de um padrão para dados médicos amplamente utilizados atualmente, pela flexibilidade para representação de informações, o uso de terminologia controlada, e a possibilidade de embutir outros objetos no padrão como imagens e eletrocardiogramas. Agregam-se desta forma, as inúmeras vantagens da utilização de padrões para registros clínicos aos dados dos sistemas antigos.

Quadro 8: Exemplo de Análise Laudo Alemão

NNH-CT		
		Snomed
WO	NNH	T22000
POSITON	Bereich	
WAS	Sekretspiegel	M02590
WIE-WAS	Kein	
WAS	Pnematisation	F75000
WIE	Regelrechte	
WAS	Sinusitis	M40001
WIE	Kein	
Thorax-CT		
WO	Thorakalskelett	T10300
WIE	Regelrechts	
WO	Trachea	T25000
WIE	Mitelständige	
WO	Lunge	T28000
WIE	frei	
WAS	Pneumonie	M40000
WIE	kein	
WO	Herz	T32000
WIE	Mitelständige	

Fonte: Acervo do Autor

Ao observar-se o Quadro 8, verifica-se a mesma estrutura das informações chaves, realizando-se a tradução: O que, Como Onde, Posição (Was,Wie,Wo Position).Abaixo Tela Geral do Sistema de Análise Laudos Médicos.

Figura 22– Tela Principal do Sistema

Fonte: Acervo do Autor

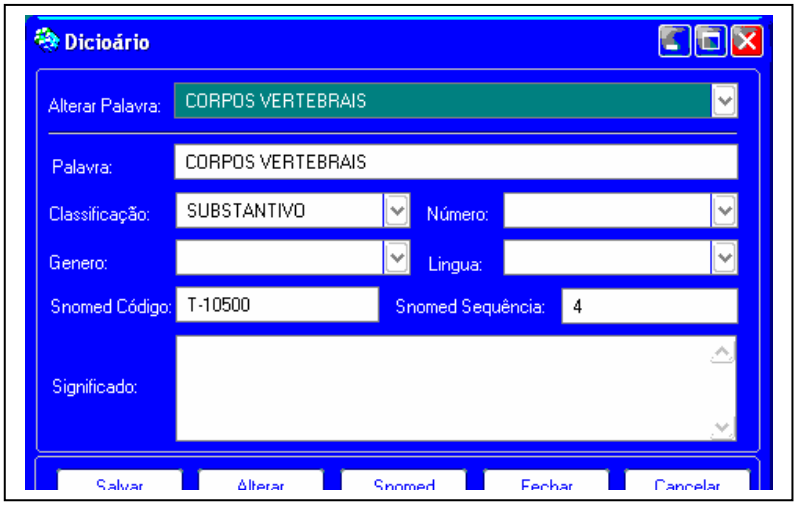


Figura 23 – Exemplo de Dicionário Palavras

Fonte: Acervo do Autor

Em seguida, um exemplo da Interface de Análise Laudos Médicos Coluna Vertebral no Idioma Português.

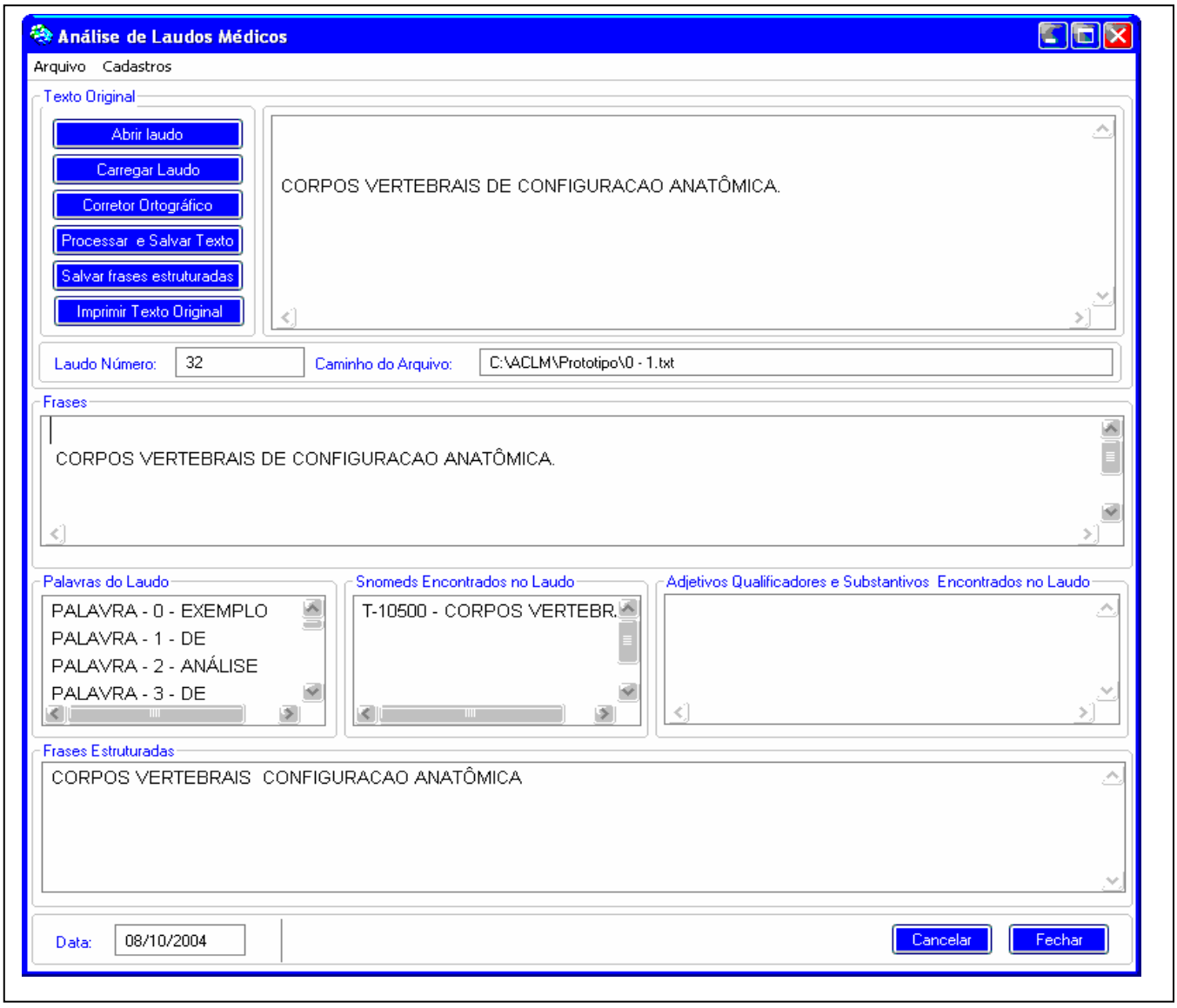


Figura 24– Exemplo de Análise Laudo em Português

Fonte: Acervo do Autor

E a seguir, na próxima figura um laudo analisado da área do tórax, que possui mais frases e estruturas anatômicas do que a anterior.

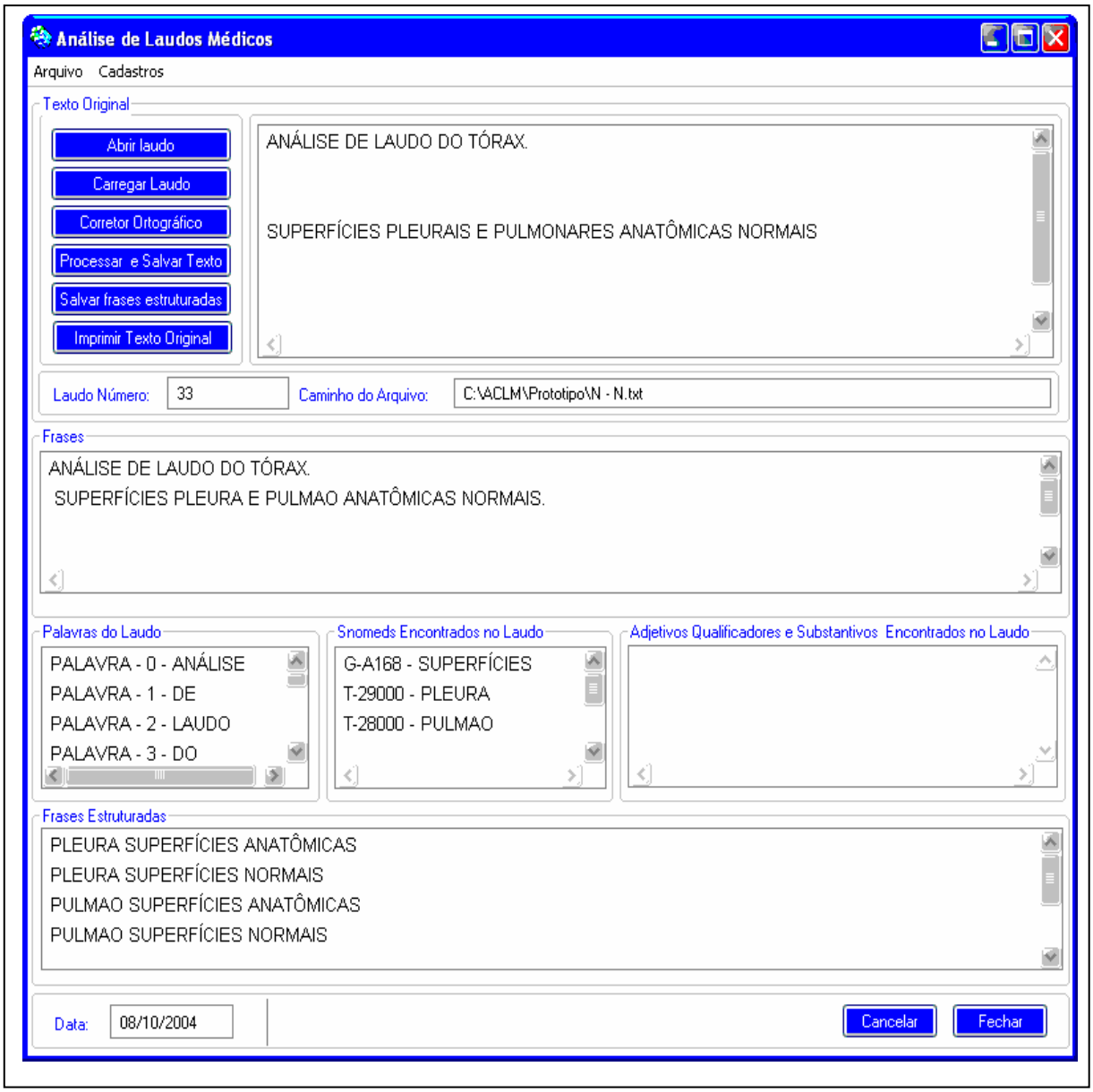


Figura 25 – Exemplo de Análise Laudo

Fonte: Acervo do Autor

A próxima figura apresenta o resultado da análise após a geração do arquivo de exportação para o padrão DICOM-SR, já devidamente interpretado pelo sistema que gera laudos estruturados.

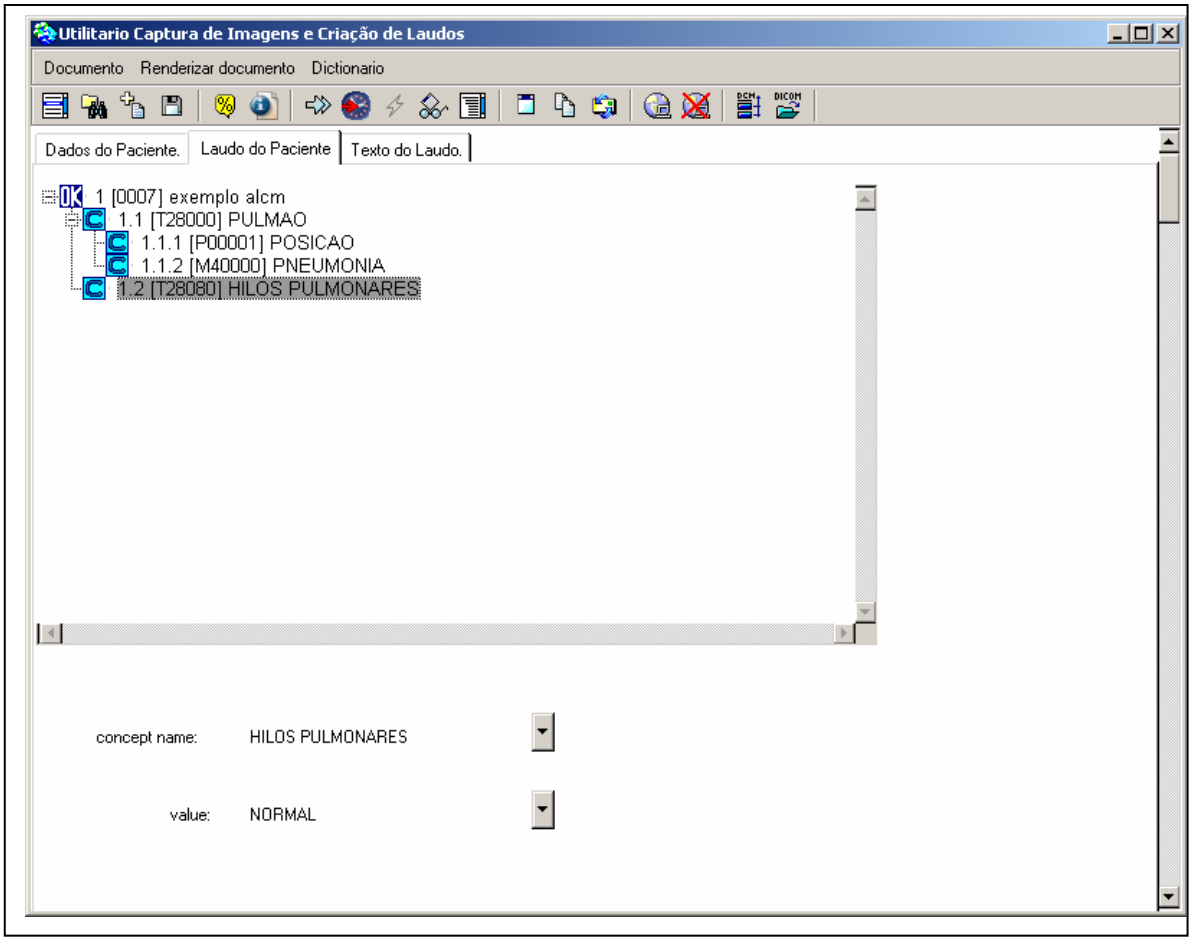


Figura 26– Exemplo de Análise Laudo DICOM SR em Alemão

Fonte: Acervo do Autor

O próximo exemplo está no idioma Alemão, também com a imagem das telas do software de análise e do software de DICOM-SR.

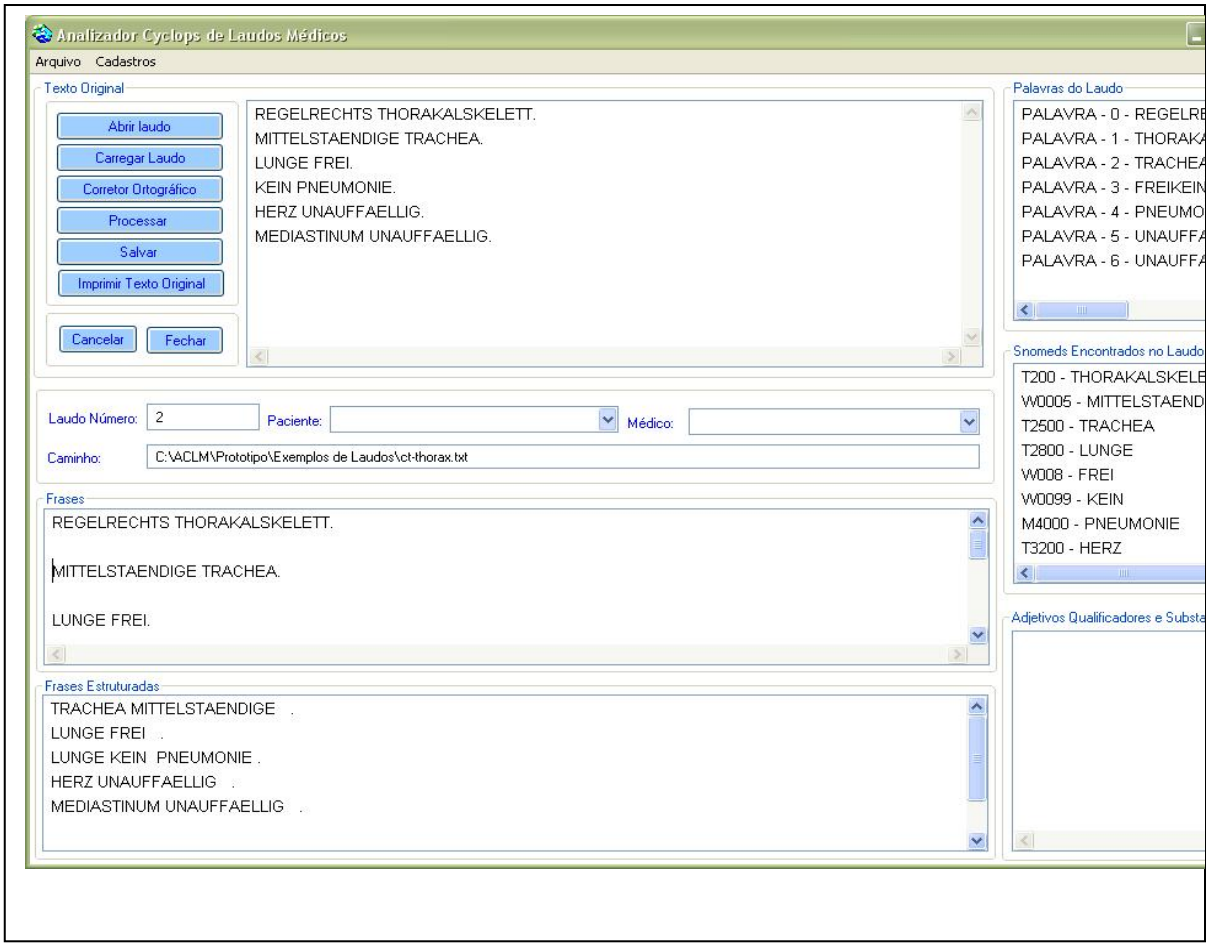


Figura 27– Exemplo de Análise Laudo Dicom SR em Alemão

Fonte: Acervo do Autor

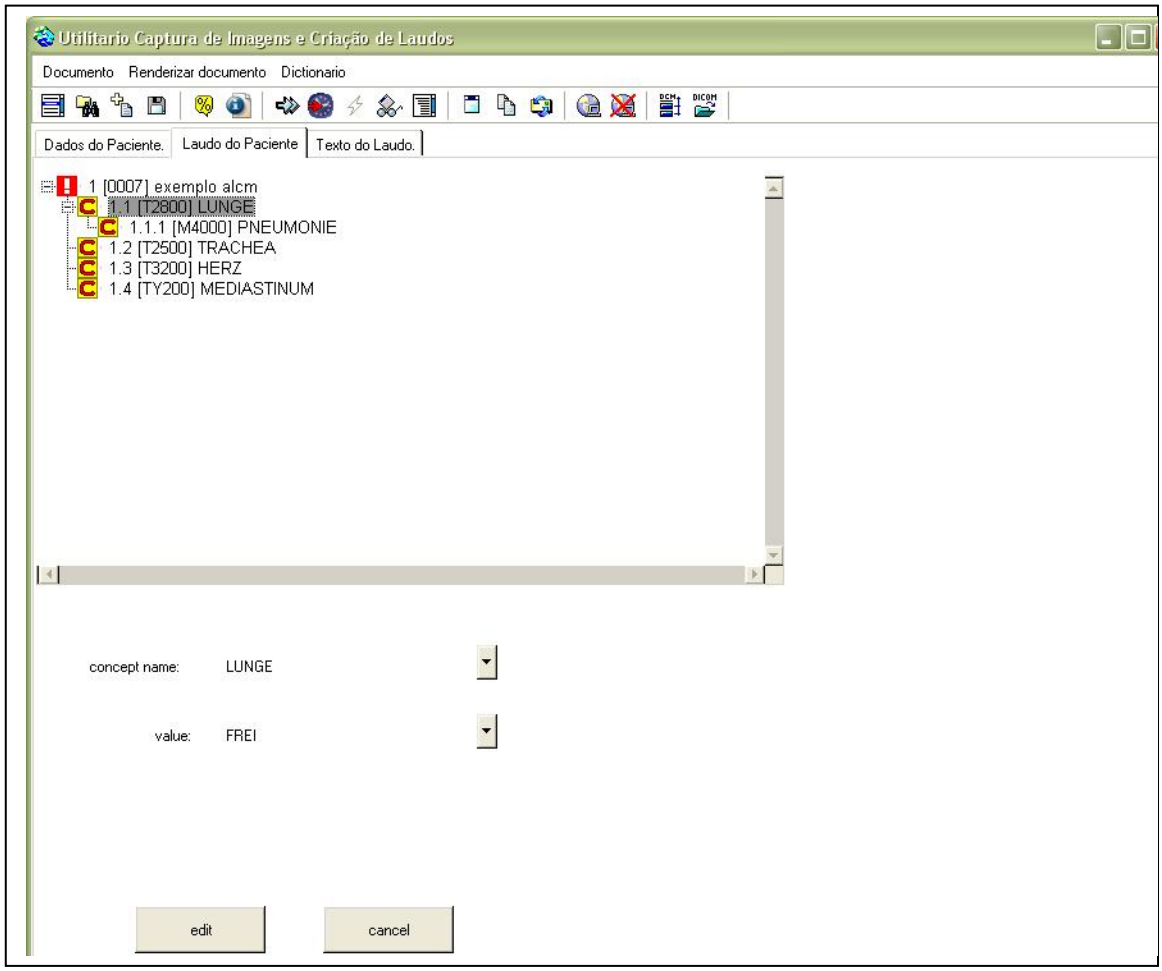


Figura 28– Exemplo de Análise Laudo DICOM SR em Alemão

Fonte: Acervo do Autor

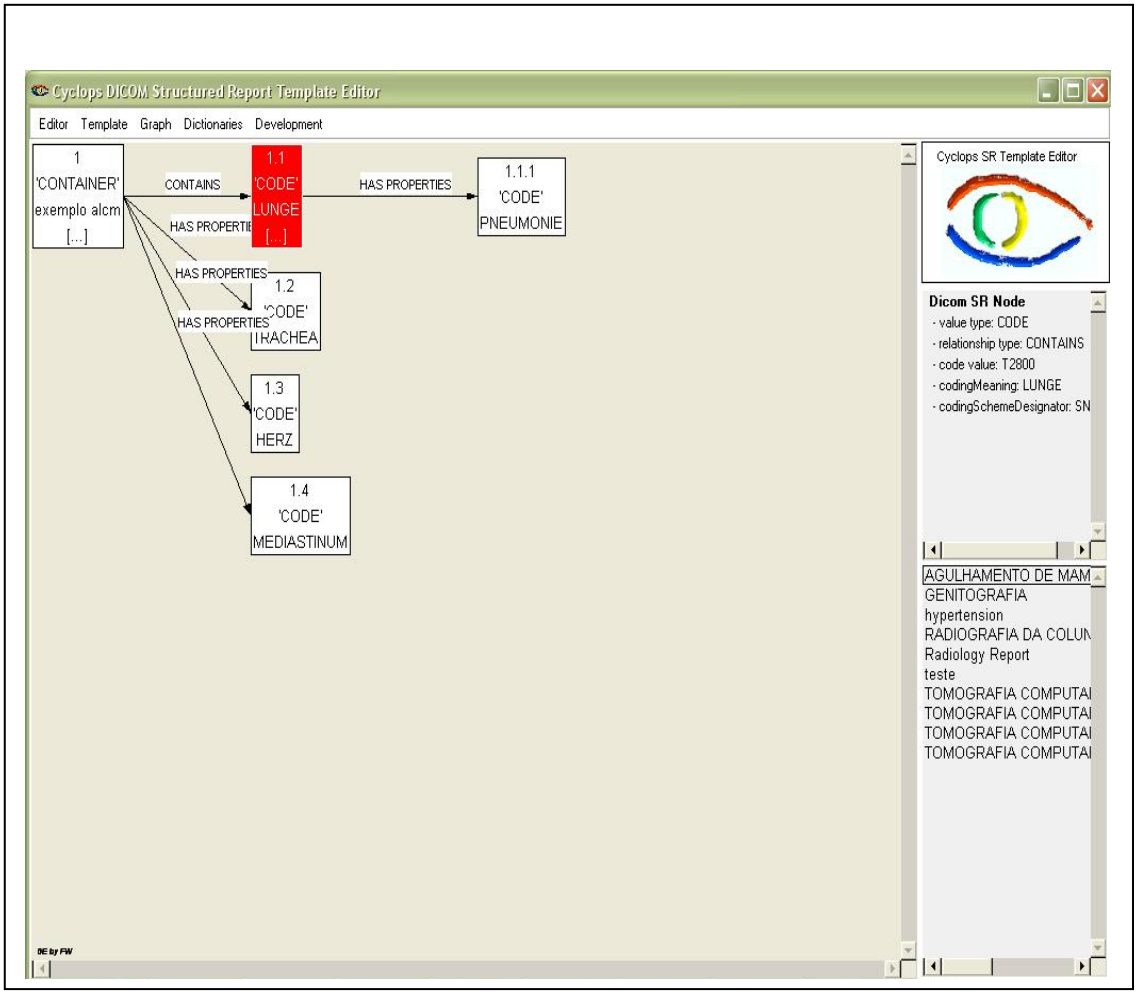


Figura 29– Exemplo de Análise Laudo DICOM SR em Alemão

Fonte: Acervo do Autor

7 Considerações Finais

O Modelo de Recuperação e Interpretação de Laudos, utilizando as palavras contendo as informações mais importantes a serem extraídas como a região Anatomia (onde?) as Doenças ou alterações identificáveis (o que?) e a qualificação de cada uma delas (como?) e Como (do o que)?), além da identificação da posição, mostrou bons resultados. Porém uma das maiores dificuldades encontradas, sendo proibitiva para mais testes, foi à ausência de Nomenclatura SNOMED na língua portuguesa por completo e versão atualizada na língua alemã.

Pois o custo, mesmo para uma licença de uso, com finalidade apenas para a pesquisa, não sendo permitido a aplicada em outras instituições como hospitais, clínicas e outros, é de seis mil dólares, restringindo assim os testes realizados e sendo necessária a criação de codificação e classificação das demais palavras não constantes na versão disponível no livro em alemão de 1984, bem como a tradução para as existentes.

O protótipo de software está tratando diversos textos após a sua conversão ao formato txt padrão ASCII, sofrendo a correção ortográfica e também foram realizadas as analises léxicas, sintáticas, morfológicas e de discursos das frases para formação de estruturas de informações encontradas. Para integração aos softwares do Projeto Cyclops que utilizam o padrão DICOM SR, foram desenvolvidas rotinas de geração de código XML que são importados pelos mesmos.

Também, referente ao tema deste trabalho, foi aceito um artigo no congresso internacional Current Aspects of Knowledge Management in Medicine (KMM05) na Alemanha, que se encontra em Anexo.

7.1 Recomendações de Trabalhos Futuros

Aplicar o protótipo em ambiente de produção hospitalar, visando ao seu aprimoramento e torná-lo uma ferramenta de trabalho disponível às instituições hospitalares.

Continuidade da referida pesquisa sobre o tema expandindo para outros temas dentro da área médica hospitalar.

Formação de um comitê mutildisciplinar junto ao CFM – Conselho Federal de Medicina e ao Colégio Americano de Patologias, para apoiar e suportar o uso da Nomenclatura SNOMED no Brasil traduzindo-a para o idioma Português.

8 Referências Bibliográficas

[ABID e MANICKAM2002] SSR ABIDI & S Manickam. **Extracting Case Structures from XML-Based Electronic Patient Records**: A Knowledge Engineering Solution to Augment Case Based Reasoning Systems. International Journal of Medical Informatics, 2002.

[BAKKEN E .WARREM2002] BAKKEN, S.,, WARREN J. J. **An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED Clinical Terms®**, International Journal of Medical Informatics, 2002.

[BELIAN2002] BELIAN, Rosalie Barreto; Novaes, Magdala de Araújo. **Tópicos Relevantes no Desenvolvimento do Prontuário Eletrônico do Paciente,** CBIS'2000 - Congresso Brasileiro de Informática em Saúde. São Paulo, 2000.

[BELLAZI e MONTAINI 2001] BELLAZI, R., MONTAINI S. **Cased-Based Reasoning for medical knowledge-based systems.** Int J Med Inf. 2001.

[BOTOLUZZI2002] BOTOLUZZI, Mariana K, **Desenvolvimento de um Editor de Laudos Compativel com o Padrão Internacional DICOM Structured Reporting**; VIII Congresso Brasileiro de Informática em Saúde. Sociedade Brasileira de Informática em Saúde, 2002.

[BROW e WARMINGTON 2002] BROWN Philip J. B., WARMINGTON Victoria, **Data quality probes—exploiting and improving the quality of electronic patient record data and patient care.** International Journal of Medical Informatics, , 2002,

[CHRISTODULAKES2000] CHRISTODULAKES, Demtres - **Natural language processing**: 2th International Conforerence, Patras Greece: proceedings NLP - BERLIM - Springer, 2001.

[CLUNIE2000] CLUNIE David A. - **Natural DICOM Structured Reporting**, PixelMed Publishing, Bangor, Pennsylvania, 2000.

[DICOM2003] **DICOM Digital Imaging and Communications in Medicine.** [On-Line] available: http://medical.nema.org/

[ELKIN E BROWN2002] Elkin PL, Brown SH, **Guideline and quality indicators for development, purchase and use of controlled health vocabularies**. Int J Med Inform. 2002.

[FALLER1999] FALLER, Adolf - **Der Körper des Menschen**: Einführung in Bau und Funktion - STUTTGART - Thieme, 1999.

[GNU2003] GNU, **Aspell**. Disponível em < http://www.gnu.org/software/aspell>.Acesso m: 7 maio 2003.

[JACKSON2002], JACKSON, Peter**, Natural Processing for online applications**:, J.Benjamins Pulishing Co., Philadelphia, 2002.

[HL72004] **HL7 Health Level Seven**, American National Standards Institute (ANSI) . [On-Line] available: http://www.hl7.org.

[LIU1996] LIU, Li-min, **Modeling a vocabulary in an object-oriented database**, Conference on Information and Knowledge Management archive. Rockville, Maryland, 1996.

[MEDEIROS1998]. MEDEIROS, Aderaldo, **A Língua Portuguesa** [On-Line] available: http://www.linguaportuguesa.ufrn.br;

[MÖLLER1998] MÖLLER, Torsten B, REIF, Emil, **CT – Und MRT – Normalebefunde**, Stuttgart, Thieme, 1998.

[MRE1996]. Ministério das Relações Exteriores – Governo do Brasil, CPLP - **Comunidade dos Países de Língua Portuguesa**.[On-Line] available: http://www.mre.gov.br/cdbrasil/itamaraty/web/port/relext/mre/cplp/apresent.htm.

[PHILLIPS E BUCHANAN2001] Phillips, J., Buchanan, B.G., **Ontology-guided knowledge discovery in databases.** International Conf. Knowledge Capture Victoria, Canada, 2001.

[RUSSEL & NORVIG1995**], Artificial Inteligence**, Pearson Education Inc. 1995.

[SNOMED2004] **SNOMED Systematized Nomenclature of Medicine**. SNOMED International, a division of the College of American Pathologists (CAP). [On-Line] available: http://www.snomed.org.

[STEENKISTE E JACOBS 2001] van STEENKISTE BC, JACOBS JE, **A Delphi technique as a method for selecting the content of an electronic patient record for asthma**. Int J Med Inf. 2001

[THEWS1999] THEWS, Gerhard - **Pathophysiologie des Menschen** - STUTTGART - Wiss.Verl.Ges, 1999.

WANGENHEIM2003] Von Wangenheim, Aldo, von Wangenheim G. Christiane, 2003, **Roteiro de Engenharia de Software para Pesquisadores do Projeto Cyclops/Brasil**, não publicado, 59p.

[WANGENHEIM1993] von Wangenheim, Aldo, 1993, **Uma Interface em Linguagem Natural de Aplicação Genérica:** Estudo de Viabilidade e Implementação, não publicado.

[WILLIAMSON1995] WILLIAMSON, Robert E. **ANNOD: a navigator of natural-language** organized (textual) data, Annual ACM Conference on Research and Development in Information Retrieval archive, Montreal, Quebec, 1995.

[WINGERT1994] WINGERT, Friedrich, **SNOMED Systematisierte Nomenklatur der Medizin,** Heidelberg: Springer-Verlag, 1984.

[ZETKIN1980] ZETKIN, Maxin **- Wörterbuch der Medizin** - BERLIM - Verl.Gesundheit, 1980.

ANEXO

ANEXO: Artigo KM2005 Kaiserslautern Alemanha.

Improving Content Based Recovery on a Radiological Reports Database

FÁBIO ALEXANDRINI[1], MARIANA KESSLER BORTOLUZZI[2], ALDO VON WANGENHEIM[3]

[1,2,3] *The Cyclops Project Department of Computer Science Federal University of Santa Catarina: 88049-900 Florianópolis, SC, Brazil*

[2] *Department of Business Information Systems II University of Trier 54286 Trier, Germany*

[1]fabalex@unidavi.edu.br; [2]kesslerb@uni-trier.de; [3]awangenh@inf.ufsc.br.

Abstract. The present effort focuses on developing a method for assisting the representation of radiological reports, written in simplified natural language, in a standardized and content recovery prone structure, such as DICOM Structured Report. Sample reports were collected and have being analyzed. The work is currently in process, but an intermediary representation was already reached and is being evaluated by physicians to attest the accuracy of the results.

1 Introduction

Most health care institutions have a precious legacy of clinical reports written in natural language, or simplified grammatical structure. Unfortunately content based retrieval of information from these reports is inefficient due to the peculiarities of natural language. This prevents institutions from sharing clinical records without waste of precious time and resources.

The present research effort focuses on the development of methods based on knowledge about normal findings in radiological examinations [1] and the Systematized Nomenclature of Medicine - SNOMED [2] with the objective of translating thoracic radiological reports into a representation more suitable for content recovery and that can be, in further work, rendered into reports compliant to internationally accepted standards such as DICOM Structured Report [3]. A set of radiological reports, provided by a Brazilian and a German health care institution, was used as source of sample subjects for the experiment.

2 State of the art

The natural language processing involves the development of intelligent computer systems that deals with problems in microworlds, application limited domains characterized by the search of most appropriated technique to solve each sort of problem. It seeks to develop systems capable to solve complex problems composed by distinguished tasks [4]. Some subfields of infomation retieval rely on a training corpus of documents that have been classified as either relevant or non-relevant to a particular situation, in text categorization or attempts to assign documents to two or more pre-defined categories [5].

In the same way there must be a more appropriated technique to solve each task, where several tools of AI and other areas of knowledge are combined in only one intelligent system that manages the efforts to solve tasks.

Clinical report, especially in radiology contains information concerning a patient's medical condition. However, a great percentage of this information is not structured, for it's free text based, that consequently makes it more difficult to search, analyze, summarize and present.

Previous studies have shown the potential benefits of medical structured data for the practice, research and medical teaching. The information's structure can be used to help organize and improve the medical record presentation [6], [7], [8], [9], [10].
Specialist systems can use the structured information of Clinical report to decision support [11], [12], [13]. For research and teaching, structured Clinical Reports can extremely improve recall and the precision of recovering information's tasks. Only structured data are accessible to the cause, space, time and evolutionary advanced database that models the technique being developed on computing and medical informatics fields [14], [15], [16].
Other systems that seek to enlarge the semantic contents of ontology to guide the knowledge discovery on database and also analyze the variables types that the user checks [17]. Systems that work on evaluation and comparison of terminology models use of diagnosis concepts for clinical terms terminology for integration, as SNOMED, have obtained success with semantics categories of ECS – European Committee for Standardization of structured categories and ISO reference model of terminology [18]. But there are also problems concerning information quality, one of the main factors of successfully or unsuccessfully cases of application or methodologies proposed.
One fact that deserves to be singled out is the lack of works directed to applications in the Portuguese Language, face to the fact that most terminologies are found mostly in English, having a few versions to other languages as German, French and Spanish. Portuguese is the eighth most spoken language on the planet, third among occidental languages, after English and Spanish [19].

3 The DICOM SR Standard and the SNOMED Nomenclature

SNOMED is a widely accepted terminology and infrastructure designed to enable the sharing of health care knowledge, across clinical specialties and sites of care. It contains, preferred medical terms and concepts, consisting of more than 144,000 terms and term codes divided into eleven: Topography; Morphology; Function; Living Organisms; Chemicals, Drugs, and Biological Products; Physical Agents, Activities and Forces; Social Context; Diseases/Diagnoses; Procedures; and general Linkage Modifiers. SNOMED is available in several languages including German. Unfortunately there is not yet an available translation of the SNOMED terms for Portuguese. Thus, in order to make this experiment possible, the basic terms for thoracic radiological exams were translated by collaborator Brazilian physicians.

The DICOM SR standard sets out rules that define how structured documents that contain health information should be composed, stored and transmitted. These make use of a controlled terminology; which enhances the results of content based retrieval [20].

4 Methods

The sample thoracic reports, 315 of them in German and 7719 in Portuguese, were analyzed. In the sample reports a common structure was identified. A heading containing the identification information such as of the physician, patient, and institution in which the report was emitted, followed by a description of the techniques used to perform the exam, such as type of procedure. The main part of the report is the body in which the conclusions drawn from the exam can be found. This will be the focus of the effort to interpret, retrieve and codify content.

The language used by physicians to describe findings in radiological examinations in both the institutions can not be considered indeed natural language. It is rather a set of short sentences normally without the use of verb, escaping the linguistic conventional rules. These, often either affirm of deny the existence of a malformation or alteration in an anatomical structure, the presence of strange bodies or illnesses, adding observations regarding the localization, shape or appearance of the anatomical structure. For each type of clinical report of radiological examination there are specific anatomical structures of interest [1].
The subject of most sentences found in thorax radiological reports are anatomical structures such as *mediastinum, lung, trachea* and others. The subject is often followed by adjectives stating the position, or the part of the organ that was observed, for instance *surface* or *right*. Adverbs, describing morphological information follow, such as *normal* or *reduced*. To this, might follow information regarding diseases, like for example, the existence of emphysema and adverbs indicating the degree of severity of the disease. For the analysis and structuring process ware used natural language processing techniques, lexical, syntactic, semantics and of speech analysis combined with information from the specialist physicians. Because the medical language and usually language have a great difference, normality the language in the clinical reports are affirmative or negative phrases with few or without verb.

5 Structuring Approach

The sample radiological reports from Dataflex and word documents base were converted to simple text ASCII standard representation in order to disregard formatting information. Afterwards, the GNU Aspell free tool [21], available in several languages, was used to perform a lexical checking.
For the natural language processing, the text file is separated in sentences observing punctuation and end of line commands. Later the words are separated for the lexical analysis. Using a former developed dictionary, the words are classified by their grammatical class, such as substantive, adjective and others. The sentences and words are the input for the syntactic and semantics analysis. Each word has a function in the sentence, but we search first the terms for anatomical structures, diseases and morphology are then matched to the appropriate SNOMED codes.
After the first tests of classification, there was found a great amount of adjectives that derived from substantives and for that reason had a huge importance, needing then a special treatment, for were directly referred to anatomies. See the table 1. For each sentence without an anatomical SNOMED matching term, a speech analysis is used, the algorithm searches an anatomical term from the previous phrases. When no term is found, the code of anatomical term from the clinical report's title is used.

Table 1. Example of word classification

Word			grammatical class	Snomed	Refe-rence
Portuguese	English	Germany			
Pulmão	Lung	Lunge	substantive	T28000	
pulmonar	pulmonary	pulmonal	adjective	T28000	Lung
Pleura	Pleura	Brustfell	substantive	T29000	
pleural	pleural	pleural	adjective	T29000	Pleura
normal	normal	normal	adjective		
Alteração	Alteration	Änderung	substantive		
anatômico	anatomic	anatomisch	adjective		

Earlier the use of specific rules for each type of exam was used, but due to the complexity and difficulty to make a great number of rules, computationally were studied the resemblance and differences among the medical reports and the sentence structures contained on those. Despite the diversity of sentences as well as in numbers of words as in the way of position or construction, the most important information to be extracted are the anatomical region (where?), the diseases or alterations (what?), and the qualifications of those (how?). Additionally can be used the position into anatomical region and general modifiers, such as *right, left, bilateral, surface, superficial* and others.

The terms for anatomic regions are searched in the "Topography" branch of SNOMED. If a word regarding position is along with the anatomic region, that also can be converted to one of the terms in the "General Linkage and Modifiers" branch of the SNOMED hierarchy, and the diseases or alterations (what?) can be converted in categories "Morphology", "Diseases/Diagnoses" or "Function". In this part of the process we obtained the support from the Teachers and Resident Doctors in General Surgery from HRAV (Hospital Regional Alto Vale).

In association with the overall are found the terms of qualification (how?) such as: *normal, intact, anatomic*, etc, and can also be combined with the denial as: *without, absent*, etc, that indicate the absence of problems, or their intensity such as: *minimum, light, acute, chronic*, etc. In the following example of analysis, on the first original sentence of the chart, there are three anatomic regions and one qualification, that would generate three sentences according to the where and how items, in a relation N anatomies to 1 qualifier, and the word "*bronchi*" also applies to "*lobar*".

On the second sentence is noticed the relation of 1 to 1 and on the last sentence is found a relation of 1 anatomy to N qualifiers, besides finding the pleural derived object that needs to be converted for the substantive *Pleura* added by the adjective *surface*, that can be found respectively in the "Topography" and "General Linkage Modifiers" branches of SNOMED. For each anatomical SNOMED term and adjective in a phrase, a new sentence must be created; this process can be visualized on the Table 2.

Table 2. Example of sentence analysis

Original Sentence	Where		How	Sentences Generated
	Anatomy (Substantives)	Snomed	Qualification (Adjectives)	
Trachea, Main and Lobar Bronchi, of normal gauge.	Trachea	T25000	Gauge Normal	3
	Main Bronchi	T26000		
	Lobar Bronchi	T26200		
Pulmonary Anatomic Hilo	Pulmonary Hilo	T28080	Anatomic	1
Pleural anatomic surface without alteration	Pleura Surface = **Position**	T29000 G-A168	Anatomic	2

6 Conclusion

The language used by physicians to describe findings in radiological examinations in both collaborator hospitals is a simplified restricted one. Nevertheless, there is much research still to be done before a reliable representation of such reports in a standard such as DICOM SR can be obtained.

The objective of this work is not to produce software capable of performing a complete automatic translation of clinical reports written in natural language to a standardized form, but to develop an approach to facilitate de process of structuring documents, so that these are better suited for content based retrieval.
Although it is currently being tested using reports written in Portuguese and German, the same method slightly adapted is expected to function with other languages. The results so far reached are at the present time subject to the evaluation of the physicians to attest the validity and suggest better approaches.

Acknowledgements

The authors thank the physicians of the Chirurgic Residence of the Upper Itajaí Valley Regional Hospital in Brazil for the translation of Radiological SNOMED terms to Portuguese and for the anonymized sample reports provided for the analysis. The Buddenbrock Blasinger und Benz Radiological Clinic, in Mainz, Germany, contributed providing anonymized sample reports. We also thank the German Academic Exchange Service -DAAD- for the scholarship number A/03/42304 granted to one of the authors.

References

1. Möller, Torsten B.: Normal Findings in Radiology. Georg Thieme Verlag, 2000

2. College of American Pathologists: SNOMED - Systematized Nomenclature of Medicine. College of American Pathologists, 1994

3. NEMA.: Digital Imaging and Communications in Medicine (DICOM): Version 3.0; 2000

4. Russel, S., Norvig, P., Artificial Intelligence - A Modern Approach. Pearson Education Inc, 1995

5. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing 6.ed. MIT. Massachusetts, 2003

6. Langlotz, Curtis P., Automatic Structuring of Radiology Reports: Harbinger of a Second Information Revolution in Radiology, Radiology, 2002

7. Hripcsak, G. et al.: Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. Radiology, 2002

8. Ricky, K. T., Stephen, G., Soderland, R. M. J.: Automatic Structuring of Radiology Free-Text Reports. Radiology, 2001.

9. Shortliffe, E.H., Hubbard, S.M.: Information systems in oncology. In: De Vita VT, Hellman S, Rosenberg S, eds. Cancer: principles and practice of oncology. Philadelphia, Pa: Lippincott, 1989

10. Aberle, D.R., et al.: Integrated multimedia timeline of medical images and data for thoracic oncology patients. RadioGraphics, 1996

11. Lyman, M, S. N., et al.: The application of natural-language processing to healthcare quality assessment. Med Decis Making, 1991

12. Brown, P. J. B., Warmington, V.: Data quality probes—exploiting and improving the quality of electronic patient record data and patient care. International Journal of Medical Informatics, 2002

13. Sager, N., et al.: Medical language processing: applications to patient data representation and automatic encoding. Methods Inf Med, 34:140–146, 1995

14. Muller, R., et al.: A graph-grammar approach to represent causal, temporal and other contexts in an oncological patient record. Methods Inf Med, 35:127–141, 1996

15. Abidi. R., Manickan, S.: Extracting Case Structures from XML-Based Electronic Patient Records: A Knowledge Engineering Solution to Augment Case Based Reasoning Systems. International Journal of Medical Informatics, 2002

16. Jackson, P.: Natural Processing for online applications. J.Benjamins Pulishing Co., Philadelphia, 2002

17. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. International Conf. Knowledge Capture Victoria, Canada, 2001

18. Bakken, S., Warren, J. J.: An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED Clinical Terms ®, International Journal of Medical Informatics, 2002

19. Medeiros, A.: A Língua Portuguesa [On-Line] available online URL: http://www.linguaportuguesa.ufrn.br, 2004

20. Clunie, David A.: DICOM Structured Reporting. PixelMed Publishing, 2000

21. Atkinson, K.: GNU Aspell. Available online URL: http://aspell.net/, 2004

GLOSÁRIO

Sigla	Descrição	Significado
SNOMED	Systematized Nomeclature of Medicine	Sistema de codificação de termos usados na medicina que compreende, doenças, parte do corpo, vírus, etc...
HL7	Health Level Seven	Nome norma ANSI que estabelece os padrões para a troca, integração, compartilhamento e recuperação de informações eletrônicas relacionadas à saúde
DICOM	Digital Imaging and Communication in Medicine	Padrão para arquivamento comunicação e gerenciamento de informações médicas.
DICOM SR	DICOM Structured Report	Parte do padrão DICOM relativa ao armazenamento de informações estruturadas como Prontuários, Laudos, etc....
CID-10	Classificação Internacional de Doenças (ICD-10)	Classificação de todas as doenças identificadas sua utilização é obrigatório em diagnósticos.
CFM	Conselho federal de medicina	órgão que possui atribuições constitucionais de fiscalização e normatização da prática médica.