UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

FLÁVIO BARBOSA

Metodologia para estruturação de informações de laudos radiológicos

Ribeirão Preto – SP
- 2013 -

FLÁVIO BARBOSA

Metodologia para estruturação de informações de laudos radiológicos

Tese a ser apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Doutor em Ciências Médicas.

Área de Concentração: Clínica Médica - Investigação Biomédica.

Orientador: Prof. Dr. Valdair Francisco Muglia

Co-orientadora: Profa Dra. Maria Cristiane Barbosa Galvão

Ribeirão Preto – SP
- 2013 -

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

FICHA CATALOGRÁFICA

Barbosa, Flávio
Metodologia para estruturação de informações de laudos radiológicos/ Flávio Barbosa; orientador: Prof. Dr. Valdair Francisco Muglia – Ribeirão Preto, 2013.
164. f. : il.

Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. Departamento de Clínica Médica. Área de Concentração: Investigação Biomédica.

1) Laudo radiológico. 2) texto livre. 3) estruturação da informação. 4) ontologia. 5) terminologia.

FOLHA DE APROVAÇÃO

BARBOSA, Flávio
Metodologia para estruturação de informações de laudos radiológicos.

Tese a ser apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Doutor em Ciências Médicas – Investigação Biomédica.

Aprovado em: ____/____/____

Banca examinadora:

Prof. Dr. _____
Instituição: _____ Assinatura: _____

Prof. Dr. _____
Instituição: _____ Assinatura: _____

Prof. Dr. _____
Instituição: _____ Assinatura: _____

Prof. Dr. _____
Instituição: _____ Assinatura: _____

Prof. Dr. _____
Instituição: _____ Assinatura: _____

DEDICATÓRIA:

À Deus:
Sempre me fortaleceu principalmente nos dias de tempestade intensa.

Aos meus genitores:
Exemplos de honestidade, humildade, resignação e bondade são meus genitores desta existência:
o Sr. Antonio Luis Barbosa, meu pai;
a Sra. Marli Luzia Barbosa, minha mãe.

Aos meus padrinhos Gaúchos:
Generosos, não tenho como exprimir minha emoção e gratidão porque encheu de possibilidades minha vida, são eles:
Excelentíssimo Dr. Luiz Gonzaga Mendes Marques e
sua esposa Sra. Tereza Noll Marques

A minha namorada, esposa e mulher:
A ela que sacrificou suas noites para cobrir minhas faltas nos afazeres familiares e que sempre me ajudou, incentivou, escutou, dialogou, amou, desejou, ai ai, etc.:
Sra. Michelle Mazer

Ao meu filho:
Ele que me inspirou a continuar lutando e compreendeu, mesmo tão novinho, quando lhe pedia para me deixar estudar:
Gabriel Gomes Barbosa

Ao meu enteado:
Respeitoso e amoroso que testemunhou minha dedicação e de sua mãe:
Leonardo Mazer Bueno da Silva

AGRADECIMENTOS:

Impossível enumerar todas as pessoas que contribuíram para que fosse possível a minha pessoa chegar a este momento. Todavia, é possível agradecer a todas àquelas que contribuíram diretamente para elaboração deste trabalho.

As instituições que prontamente atenderam as solicitações de autorização para coleta dos laudos radiológicos da coluna lombo sacra:

ao CCIFM (Centro de Ciências das Imagens e Física Médica) do HCFMRP-USP (Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo);

Cedirp;

Omnimagem.

As pessoas sem as quais o trabalho não teria alcançado o resultado esperado:

Professor Dr. Valdair Francisco Muglia:
Interrompia seus afazeres para filosofar sobre a estruturação dos dados, os fundamentos da terminologia e sempre dizia com um sorriso confiante: “Flávio, você entendeu?”

Professora Dra. Maria Cristiane Barbosa Galvão:
Transmitiu seu conhecimento ajudando a elucidar e discernir as ideias nebulosas e embaraços entre as Ciências da Informação e da Computação.

Professor Dr. Paulo Mazzoncini de Azevedo Marques:
Ajudou a focar o trabalho no que de fato era importante.

Professor Me. Wilson Moraes Goes:
Flexibilizou os horários de trabalho para permitir que os estudos pudessem acontecer com a frequência exigida.

Silvio Cesar Somera:
Discutiu questões conceituais referentes aos dados, informação e o conhecimento.

Professor Dr. Jorge Elias Junior:
Preencheu a lacuna do que os radiologistas e solicitantes precisam.

A informação como bem econômico:

"(...) Ela não precisa de porto, estrada e nem de avião porque são transportes muito lentos para ela. Desconhece as fronteiras físicas e transita em uma velocidade parecida com a da luz. Não tem passaporte e não reconhece os muros! Por que você acha que aquele muro foi derrubado? Porque não barrava o que tem importância hoje, ele ficou inútil.

E para desespero de todas as leis econômicas, esta senhorita informação, não admite ser trocada por nada, só topa ser partilhada. Quando eu compro o seu carro você fica sem ele, mas quando eu compro uma informação de você nós dois ficamos com ela e você ainda por cima com o meu dinheiro, entendeu? Porque aquele cara é o homem mais rico do mundo? (...)"

Valdez Ludwig

RESUMO

BARBOSA, F. **(Metodologia para estruturação de informações de laudos radiológicos**). 2013. 164 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2013.

Introdução: O laudo radiológico com entrada de dados estruturada proporciona precisão na recuperação de informações, todavia, é menos abrangente que o texto-livre e pode impedir o especialista de descrever um achado importante, não contemplado pela estrutura pré-definida. Portanto, a solução está em interpretar o texto escrito pelo especialista e os transportar para uma estrutura computacional ontológica formalizando o conhecimento.
Objetivos: Elaborar e avaliar uma metodologia para estruturação de informações de laudos radiológicos em texto livre que contemple as especificidades brasileiras, incluindo a terminologia empregada.
Métodos: Coleta de 1701 laudos radiológicos de ressonância magnética da coluna lombossacra em três instituições distintas. Para estruturar a informação, foram utilizadas técnicas de mineração de texto e conceituação ontológica das unidades lexicais. Dez radiologistas especialistas avaliaram a superestrutura textual e a terminologia extraída e conceituada, por meio de questionário eletrônico onde foram sugeridas 3 superestruturas textuais e 30 termos com a descrição do respectivo conceito.
Resultados: A metodologia criada é constituída por 6 passos: 1) coleta dos laudos radiológicos de determinado tipo de exame; 2) decomposição textual; 3) normalização das unidades lexicais; 4) identificação das superestruturas textuais; 5) conceituação dos termos candidatos; 6) avaliação das superestruturas identificadas e da terminologia extraída e conceituada pelos especialistas na linguagem de especialidade estudada. Foram identificadas três superestruturas textuais diferentes. Após processamento automático inicial, 6129 unidades lexicais foram normalizadas obtendo 4568 termos candidatos dos quais 4183 foram conceituados e originaram 727 conceitos. Foram estabelecidas 13963 relações entre termos candidatos e conceitos, e 789 relações entre conceitos. Quanto a adequada descrição dos termos, 40% dos avaliadores concordam plenamente que estão adequados, 60% concordam parcialmente que estão adequados e nenhum avaliador discordou que os termos são adequados. Em relação a quantidade de termos utilizados pelos avaliadores, em sua respectiva instituição, 70% utilizam de 24 a 30 termos e 30% de 16 a 23 termos. A conceituação de um termo foi considerada inadequada por dois avaliadores.
Conclusão: A metodologia proposta foi avaliada por radiologistas especialistas que identificaram as diferenças das superestruturas textuais entre as instituições e os termos com a descrição do respectivo conceito. Portanto, foi possível estruturar a informação de laudos radiológicos da coluna lombossacra a partir do texto livre identificando as superestruturas textuais, extraindo as unidades lexicais, normalizando as unidades lexicais e conceituando ontologicamente os termos

candidatos preservando a referência com as respectivas partes constituintes dos laudos radiológicos produzidos no contexto brasileiro.

Descritores: Laudo radiológico, texto livre, estruturação da informação, ontologia, terminologia.

ABSTRACT

BARBOSA, F. (**Methodology for structuring information from radiological reports**). 2013. 164 f. Thesis (Doctoral) - Medical School of Ribeirão Preto city, University of São Paulo State, 2013.

Introduction: The radiological report with structured data entry provides precision in information retrieval, however, is less comprehensive than the text-free and may prevent the expert to describe an important finding, not covered by the pre-defined structure. So the solution is to interpret the text written by expert and carry for a computational framework formalizing the ontological knowledge.
Objectives: Develop and evaluate a methodology for structuring information in free text radiology reports that contemplates the Brazilian specifications, including terminology.
Methods: 1701 radiological reports of lumbosacral MRI were collected in three different institutions. To structure the information we used techniques of text mining and ontological conceptualization of lexical units extracted. Ten dedicated radiologists evaluated the textual superstructure and terminology extracted by the electronic questionnaire which were suggested textual 3 superstructures and 30 terms with the description of its concept.
Results: The established methodology consisted of six steps: 1) collection of radiology reports of certain type of examination; 2) textual decomposition; 3) lexical units normalization; 4) textual superstructures identification ; 5) conceptualization of candidate terms; 6) evaluation of identified superstructures and terminology extracted by experts in the specialized language study. We identified three different textual superstructures. After automatic processing initial 6129 lexical units were normalized getting 4568 candidate terms of which 4183 were conceptualized and originated 727 concepts. 13963 were established relations between candidate terms and concepts and 789 relations between concepts. As for adequate description of the terms, 40% of the evaluators agree fully that are appropriate, 60% somewhat agree that they are adequate and no evaluator disagreed that the terms are adequate. Regarding the number of words used by evaluators in their respective institution using 70% between 24 and 30 terms and 30% from 16 to 23 terms. The conceptualization of a term was considered inadequate by two evaluators.
Conclusions: The proposed methodology was evaluated by radiologists who identified the differences between the textual superstructures institutions and the terms describing the relevant concept. Therefore, it was possible to structure the information of the lumbosacral spine radiology reports from free text by identifying textual superstructures, extracting lexical units, normalizing the lexical units and conceptualizing the terms ontologically candidates preserving the reference to the respective constituent parts of radiology reports produced in the Brazilian context.

Uniterms: Radiological report, free-text, structuring information, ontology, terminology.

LISTA DE FIGURAS

Figura 1: Triângulo do significado 24
Figura 2: Para Teoria Geral da Terminologia um termo está associado a um único conceito. 25
Figura 3: Exemplo de superestruturas textuais em diferentes atividades especializadas 27
Figura 4: Impressão de inteligência: camadas sobre o pulso elétrico 28
Figura 5: Etapas para geração do conhecimento e inteligência (adaptado de Freire, 2012) 30
Figura 6: Macro funcionamento de um sistema de recuperação de informação 31
Figura 7: Abordagens e técnicas que fazem uso da mineração de texto (adaptado de Miner & et.al., 2012) 33
Figura 8: Processo de mineração de texto frequentemente utilizado em domínio de conhecimento especializado 34
Figura 9: Processo de mineração de texto em ciclo frequentemente utilizado em domínio de conhecimento especializado 35
Figura 10: Fragmento de um laudo radiológico da coluna lombossacra exibindo marcadores de formatação 36
Figura 11: Modelo conceitual para geração de um índice invertido 37
Figura 12: Modelo conceitual para geração de um índice invertido multidimensional 38
Figura 13: Formalização eletrônica do conceito: informação unívoca. 47
Figura 14: Exemplo de utilização da decisão ontológica contextual horizontal 49
Figura 15: Exemplo de utilização da decisão ontológica contextual vertical 50
Figura 16: O raciocínio clínico com solicitação de exame diagnóstico por imagem 53
Figura 17: Arquitetura de software da ferramenta computacional *Meta-generalis*. 62
Figura 18: Processo de anonimização na ferramenta *Meta-generalis* 65
Figura 19: Visualizando os laudos de onde o termo candidato "disrafismo" foi extraído para ser conceituado 70
Figura 20: Esquema de utilização das superestruturas textuais identificadas nas instituições 77

LISTA DE EQUAÇÕES

Equação 1: Definição do índice de densidade léxica (DL) 41
Equação 2: Definição do índice de aglutinação léxica (AL)..... 42

LISTA DE GRÁFICOS

Gráfico 1: Análise da riqueza léxica do fragmento de laudo radiológico “A” 43
Gráfico 2: Análise da riqueza léxica do fragmento de laudo radiológico “A” considerando "stopwords" 45
Gráfico 3: Resultado absoluto das normalizações por instituição 73
Gráfico 4: Densidade léxica (DL) e aglutinação léxica (AL) por instituição e tipo de normalização 74
Gráfico 5: Densidade léxica (DL) e diferença de aglutinação léxica (dAL) por instituição e tipo de normalização 75
Gráfico 6: Densidade léxica (DL) e aglutinação léxica (AL) das instituições após a última normalização.....76
Gráfico 7: Proporção de termos candidatos conceituada manualmente e automaticamente por produto terminológico eletrônico (PTE) e instituição.....79

LISTA DE TABELAS

Tabela 1: Lista dos termos candidatos extraídos do fragmento de laudo radiológico "A" 40
Tabela 2: Análise da riqueza léxica do fragmento de laudo radiológico "A" 42
Tabela 3: Lista de stopwords extraída do fragmento de laudo radiológico "A" 44
Tabela 4: Lista dos termos candidatos extraídos do fragmento de laudo radiológico "A" considerando a lista de "stopwords" 44
Tabela 5: Análise da riqueza léxica do fragmento de laudo radiológico "A" considerando lista de "stopwords". 45
Tabela 6: Ferramentas computacionais para mineração de texto e construção de ontologias. 51
Tabela 7: Descrição das amostras coletadas por instituição 59
Tabela 8: Densidade léxica (DL), aglutinação léxica (AL) e a diferença da aglutinação léxica (dAL) por instituição e tipo de normalização 74
Tabela 9: Síntese dos conceitos gerados por instituição 78
Tabela 10: Síntese das relações entre termos e conceitos 78
Tabela 11: Tabulação das respostas dos avaliadores à questão 1 (n=10). 81
Tabela 12: Tabulação das respostas dos avaliadores à questão 2 (n=10). 83
Tabela 13: Tabulação das respostas dos avaliadores à questão 3 (n=10). 83
Tabela 14: Tabulação das respostas dos avaliadores à questão 4 (n=10). 85
Tabela 15: Tabulação das respostas dos avaliadores à questão 5 (n=10). 85
Tabela 16: Tabulação das respostas dos avaliadores à questão 6 (n=10). 86
Tabela 17: Tabulação das respostas dos avaliadores à questão 7 (n=10). 86
Tabela 18: Tabulação das respostas dos avaliadores à questão 8 (n=10). 87

LISTA DE QUADROS

Quadro 1: Um parágrafo do texto Finibus Bonorum et Malorum 64

LISTA DE SIGLAS

ABNT = Associação Brasileira de Normas Técnicas.

ACR = *American college of radiology* (Colégio Americano de Radiologia).

AL = Aglutinação lexical.

BI-RADS = *Breast Imaging Reporting and Data System* (laudo de imagem da mama e sistema de dados).

BIPM = *Bureau International des Poids et Mesures* (Bureau pesos e medidas internacionais).

BVS = Biblioteca Virtual em Saúde.

CFM = Conselho Federal de Medicina.

CRM = Conselho Regional de Medicina.

CDA = *Clinical Document Architecture* (arquitetura de documento clínico).

dAL = Diferença de aglutinação lexical.

DeCS = Descritores de ciências da saúde.

DL = Densidade Léxica.

DR = Doutor.

HL7 = *Health Level Seven International* (Saúde Nível Sete Internacional).

HTTP = Hypertext Transfer Protocol (protocolo de transferência de hipertexto).

IA = Inteligência artificial.

IHTSDO= *International Health Terminology Standards Development Organization* (Organização Internacional para o Desenvolvimento de Normas de Terminologia em Saúde).

KDT = *Knonwledge discovery in text* (descoberta de conhecimento em textos).

MER = *Model entity relation* (modelo de entidade e relacionamento).

MT = Mineração de texto.

PLN = Processamento de linguagem natural.

PTE = Produto terminológico eletrônico.

RDB = *Relational database* (banco de dados relacional).

RIS = *Radiology information system* (sistema de informações radiológicas).

RSNA = *Radiological Society of North America* (sociedade radiológica da América do Norte).

SBIS = Sociedade Brasileira de Informática em Saúde.

SNOMED-CT= *Systematized Nomenclature of Medicine-Clinical Terms* (Nomenclatura Sistematizada de Medicina – Termos Clínicos).

SRES = Sistema de registro eletrônico em saúde.

TCLE = Termo de Consentimento Livre e Esclarecido.

TCT = Teoria comunicativa da terminologia.

TGT = Teoria geral da terminologia.

TI = Tecnologia da informação.

UF = Unidades federativas do Brasil.

UMLS = Unified Medical Language System (sistema de linguagem médica unificada).

XML = Extensible markup language (linguagem de marcação extensível).

SUMÁRIO

1 INTRODUÇÃO 16

1.1 TERMINOLOGIA 21

1.1.1 Linguagem especializada e comum 22

1.1.2 Unidade lexical, termo e conceito. 23

1.1.3 Reconhecendo o termo 26

1.1.4 Superestrutura e partes constituintes (categorias textuais) 27

1.1.5 Atribuindo semântica ao pulso elétrico (ciclo) 28

1.1.6 Dados, informação e conhecimento. 29

1.1.7 Anonimização 31

1.2 MINERAÇÃO DE TEXTO 32

1.2.1 Decomposição do texto em parágrafos, frases e unidades lexicais 35

1.2.2 Normalização léxica 39

1.2.3 Normalização lexical e a lista de Stopwords 43

1.3 ONTOLOGIA 46

1.3.1 Decisão ontológica contextual vertical e horizontal 48

1.3.2 Breve análise das ferramentas computacionais existentes 50

1.4 O EXAME DE DIAGNÓSTICO POR IMAGEM 51

1.4.1 A Terminologia e o laudo radiológico 54

1.4.2 Processos para elaboração de laudos radiológicos 55

2 OBJETIVOS 56

2.1 OBJETIVO GERAL 57

2.2 OBJETIVOS ESPECÍFICOS 57

3 MATERIAL E MÉTODOS 58

3.1 COMPOSIÇÃO DA AMOSTRA 59

3.2 DESENHO DO ESTUDO 59

3.2.1 Fase I: Análise terminológica 60

3.2.2 Fase II: Avaliação da terminologia 60

3.3 FERRAMENTA COMPUTACIONAL *META-GENERALIS* 62

3.4 MINERAÇÃO DE TEXTO 63

3.4.1 Aquisição dos laudos radiológicos para formação do corpus textual 63

3.4.2 Anonimizando dados pessoais 63

3.4.3 Pré-processamento de importação 66

3.4.4 Pré-processamento do texto para decomposição textual 66

3.4.5 Processamento do texto para decomposição textual 66

3.4.6 Arquivo invertido 67

3.5 CONCEITUALIZAÇÃO 67
3.5.1 Método de identificação das partes constituintes 68
3.5.2 Método de associação dos termos candidatos as suas partes constituintes de origem 68
3.5.3 Conceituando automaticamente os termos candidatos apoiados por produtos terminológicos eletrônicos 69
3.5.4 Conceituando termos candidatos manualmente 69
3.6 ASPECTOS ÉTICOS 71
4 RESULTADOS 72
4.1 QUANTITATIVO 73
4.1.1 Normalização 73
4.1.2 Superestruturas textuais identificadas 76
4.1.3 Conceitualização 78
4.2 QUALITATIVO 80
4.2.1 Sobre a superestrutura 80
4.2.2 Sobre a terminologia 84
4.3 METODOLOGIA META-GENERALIS 88
5 DISCUSSÃO 89
6 CONCLUSÃO 97
7 BIBLIOGRAFIA 99
8 APÊNDICES 109
8.1 APÊNDICE A: TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO. 110
8.2 APÊNDICE B: QUESTIONÁRIO DE AVALIAÇÃO DA SUPERESTRUTURA E TERMINOLOGIA DAS INFORMAÇÕES RADIOLÓGICAS EXPLICITADAS. 112
8.3 APÊNDICE C: DISTRIBUIÇÃO DOS TERMOS A SEREM AVALIADOS ENTRE AS INSTITUIÇÕES. 117
8.4 APÊNDICE D: MODELO DE ENTIDADE E RELACIONAMENTO (MER) DA FERRAMENTA COMPUTACIONAL *META-GENERALIS*. 118
8.5 APÊNDICE E: INTERFACES DA FERRAMENTA COMPUTACIONAL *META-GENERALIS*. 120
8.6 APÊNDICE F: TABELAS DE NORMALIZAÇÃO TEXTUAIS. 124
8.7 APÊNDICE G: LISTA DE UNIDADES LEXICAIS IGNORADAS (STOPWORD) EM PORTUGUÊS. 128
9 ANEXOS 130
9.1 ANEXO A: AUTORIZAÇÃO DE ACESSO AOS DADOS DA AMOSTRA. 131
9.2 ANEXO B: "GENERALIS: EXTRAÇÃO E CONCEITUALIZAÇÃO DE TERMOS APOIADO POR TERMINOLOGIAS E ONTOLOGIA POSICIONAL" (ARTIGO PUBLICADO E APRESENTADO ORALMENTE DURANTE O CBIS2012) 135
9.3 ANEXO C: "GENERALIS: UMA PLATAFORMA PARA EXPLICITAR O CONHECIMENTO" (ARTIGO PUBLICADO E APRESENTADO ORALMENTE DURANTE O SIC2012) 141
9.4 ANEXO D: "METODOLOGIA PARA ESTRUTURAÇÃO DE INFORMAÇÕES DE LAUDOS RADIOLÓGICOS EM PORTUGUÊS." (ARTIGO A SER PUBLICADO) 146

1 INTRODUÇÃO

Transformar dados em informação computável não é uma tarefa trivial e vem sendo discutida na área de saúde há décadas em dois eixos principais: interoperabilidade e semântica informacional.

Várias iniciativas de interoperabilidade vêm sendo criadas internacionalmente, como por exemplo, o instituto HL7 (do inglês, *Health Level Seven International*) que criou um padrão de protocolo para troca de informações clínicas o CDA (do inglês, *Clinical Document Architecture*) e dentro do Brasil, por exemplo, a Agência Nacional de Saúde complementar (ANS) que criou o protocolo padrão para troca de informações em saúde suplementar (TISS). Isso comprova a necessidade de interoperar informações entre diferentes sistemas de informação, inclusive, de instituições de saúde diferentes (Galvão & Ricarte, 2012).

A semântica informacional consiste em formalizar o dado com o seu respectivo significado, ou seja, referenciando explicitamente o objeto que ele representa no mundo real. Um exemplo disso é a Fundação OpenEHR que criou e desenvolveu o modelo dual onde é possível separar e mapear os dados e as informações através da utilização de *templates* e arquétipos (Cavalini & Cook, 2013). Além de atribuir significado ao dado, sempre que possível, adotar uma terminologia padrão torna a informação computável, neste sentido, existe o IHTSDO (do inglês, *International Health Terminology Standards Development Organization*) que foi estabelecido em 2007 com o propósito de administrar a terminologia SNOMED-CT (do inglês, *Systematized Nomenclature of Medicine-Clinical Terms*) que possui, atualmente, mais de 311 mil conceitos ativos organizados em 19 eixos (achados clínicos, artefato de registro, conceito de ligação, conceito especial, contexto social, entidade observável, escalas e estágios, espécime, estruturas do corpo, evento, força física, localização geográfica ou ambiental, objeto físico, organismo, procedimento, produto biológico ou

farmacêutico, situação com contexto explícito, substanciais e valor qualificador). Os esforços dessas organizações internacionais mostram a importância de se atribuir semântica aos dados os transformando em informação computável.

O Brasil, através da portaria número 2.073 de 31 de agosto de 2011, adotou diversos padrões de interoperabilidade e representação de informações clínicas para propiciar semântica informacional, como por exemplo, o TISS, o modelo dual OpenEHR e o SNOMED-CT. Porém, informações clínicas em texto livre fazem parte do legado das instituições de saúde e não existem muitas instituições que tenham implementado o padrão de registro eletrônico de saúde ditado pelo governo brasileiro, o modelo OpenEHR (Cavalini & Cook, 2013), sinalizando que as informações clinicas continuam a ser geradas em formatos não padronizados, como texto livre e semiestruturado, nas instituições brasileiras.

A Sociedade Brasileira de Informática em Saúde (SBIS) e o Conselho Federal de Medicina (CFM) instituíram um manual de certificação para sistemas de registro eletrônico de saúde (SRES) onde é possível encontrar um requisito, tanto na versão 3.3 de 2009 e versão 4.0 de 2011, identificado por “ESTR.02.07” baseado na norma ABNT ISO/TS número 18308 que descreve uma característica mandatória para esses sistemas: “Fazer pesquisa (de texto e dados numéricos) em todos os campos (estruturados e de texto livre)”. Portanto, é preciso existir uma forma, uma metodologia, capaz de processar o dado armazenado em texto livre ou semiestruturado e lhe atribuir semântica informacional, de tal modo, que o torne computável. Um SRES abrange todos os registros eletrônicos em saúde, portanto o sistema de informações radiológicas ─ RIS (do inglês, *Radiological Information System*) ─ também precisa possuir semântica informacional.

A busca por padrões terminológicos em radiologia é perseguida constante e incrementalmente ao longo de sua história. Vários foram os produtos terminológicos de aceitação internacional que tiveram sua origem na busca pela informação computável e a comunicação especializada, menos ambígua possível. Alguns dos produtos terminológicos criados pela comunidade radiológica são: 1) o índice ACR (American College of Radiology´s Index for Radiological Diagnoses) criado para os radiologistas classificarem as imagens médicas de seus pacientes de forma simples composto de uma primeira parte numérica para codificar a região e uma segunda para codificar a doença permitindo aproximadamente 1000 (mil) termos únicos; 2) o projeto RadLex da RSNA (Radiological Society of North America) baseado no índice ACR, porém projetado para funcionar em ambiente eletrônico, contemplava uma quantidade maior de termos únicos e um maior detalhamento; 3) o BI-RADS (Breast Imaging Reporting and Data System) (ACR, 2010) que além da terminologia propõe uma estrutura para os exames de mamografia, ultrassonografia e ressonância magnética de mama, bem definida e mundialmente, aceita inclusive no Brasil (Passman *et al.*, 2011).

Produtos terminológicos interessam a diversas comunidades e não somente àquela que gera a informação porque existe o entendimento de que a padronização terminológica traz grandes benefícios para o ensino, à pesquisa e assistência. Eles facilitam a interoperabilidade entre diferentes sistemas de informação e atribuem significado ao dado. Com esta visão, é possível entender os motivos que levaram a incorporação do projeto RadLex como um subconjunto do SNOMED-CT convergindo os esforços de diferentes comunidades de saúde para um único produto terminológico de saúde com diferentes eixos (Langlotz, 2006).

O que possibilita a construção de produtos terminológicos é o estudo dos termos escritos de forma específica em determinada área de conhecimento e requer a participação de pelo menos dois atores: o terminológo, especializado no estudo dos termos, e o especialista da área de conhecimento, pertencente à comunidade que utiliza a língua de especialidade em estudo (por exemplo, o médico, o radiologista, advogado, arquiteto, etc.). Devido à explosão de dados em suporte eletrônico, um novo ator tem trabalhado nos estudos terminológicos apoiando o terminológo e o especialista, trata-se do informata.

Devido à interdisciplinaridade deste trabalho a introdução foi dividida da seguinte forma: 1) Terminologia; 2) O exame de diagnóstico por imagem; 3) Mineração de texto; 4) Ontologias.

1.1 Terminologia

A definição de “terminologia” pode estar associada ao conjunto vocabular, conjunto de léxicos, próprio de uma ciência. Todavia, quando trata do conjunto de práticas e métodos utilizados na compilação, descrição, gestão e apresentação dos termos de uma determinada linguagem especializada é uma atividade. Ela pode, ainda, estar associada à teoria que reúne um conjunto de postulados teóricos que suportam a análise de fenômenos linguísticos concernentes à comunicação especializada.

Convenciona-se escrever “terminologia” em minúsculo quando está associada a um conjunto vocabular e “Terminologia” com a primeira letra maiúscula quando trata de atividade ou teoria. Atualmente, a Terminologia é uma disciplina autônoma que tem por objetivo o estudo do termo (Almeida, 2010; Zilio, 2012).

1.1.1 Linguagem especializada e comum

Uma língua é um sistema de sinais apropriados a uma notação (MICHAELIS, 2009). Os sinais, no universo cultural de uma comunidade linguística, equivalem ao conjunto de palavras disponíveis no idioma e formam o seu léxico. O léxico de uma língua é composto por um sistema de unidades, dito unidades lexicais, que carregam consigo o conhecimento geral desta comunidade (Reis, 2010). Portanto, a linguagem comum ou língua geral é o sistema de comunicação oral e escrita de uso cotidiano em contexto geral de uma comunidade linguística (Pavel; Nolet, 2002).

Nas atividades humanas especializadas, que exigem longo período de aprendizado para seu conhecimento e experiência para a sua prática as unidades lexicais da língua geral também são especializadas pelos indivíduos que as executam atribuindo-lhes novo significado e, criando assim, os termos de uma língua especializada. A especialização da língua ocorre porque para exercer uma atividade específica a comunicação necessita ser diferenciada, ou seja, o seu conteúdo deve ser descrito de maneira precisa, sem ambiguidades, com referência a um esquema científico determinado ou uma estrutura normativa estabelecida (Maciel, 2010).

Com a finalidade de contemplar o funcionamento das unidades lexicais especializadas, os termos, em situações comunicativas especializadas, sejam elas profissionais, acadêmicas ou científicas a Terminologia surgiu (Almeida, 2010).

1.1.2 Unidade lexical, termo e conceito.

O conjunto de palavras disponíveis em um idioma, em uma língua, é chamado de léxico (MICHAELIS, 2009). A palavra é a unidade básica do léxico na análise linguística[1] e tecnicamente é chamada de “unidade lexical” (Oliveira; Isquerdo; Alves, 2007).

Ao equacionar, mapear e identificar um determinado conhecimento o especialista o organiza mentalmente em um conceito. Portanto, o conceito é a organização mental de um conhecimento, uma unidade de pensamento ou conhecimento, é concebido através de uma operação mental, podendo ser expresso e, portanto compartilhado, na comunidade especializada através de uma unidade lexical especializada denominada “termo” (Francelin; Kobashi, 2011; Silva-Filho, 2010). Todavia, o termo não é o conceito propriamente dito, mas sim, a chave de acesso, o referencial, que leva a mente humana a recuperar a organização mental de um conhecimento (Francelin; Kobashi, 2011).

O triangulo do significado (figura 1), recuperado no trabalho de Almeida & Souza (2011), demonstra claramente que o objeto do mundo real está associado a um signo (símbolo) que, no caso deste trabalho corresponde ao termo, através de uma convenção (Almeida; Souza, 2011). A mente humana ao perceber ou experimentar o signo recupera o referencial no mundo real. Estes autores descrevem o triangulo da seguinte forma:

> *“(...) Convencionou-se, culturalmente, identificar um objeto do mundo real por um signo (aresta convenção); tal signo, quando percebido por uma pessoa, gera um conceito em sua mente (aresta percepção); o conceito é*

[1] Analise linguística refere-se ao estudo das palavras (Oliveira; Isquerdo; Alves, 2007).

formado na mente da pessoa a partir de sua vivência no mundo, por ela ter reconhecido no objeto outro objeto que já conhecia (aresta experiência). A aresta convenção é apresentada como uma linha pontilhada para explicar que um símbolo não se conecta diretamente como o seu referente (objeto do mundo), mas através de uma representação mental desse elemento no mundo."

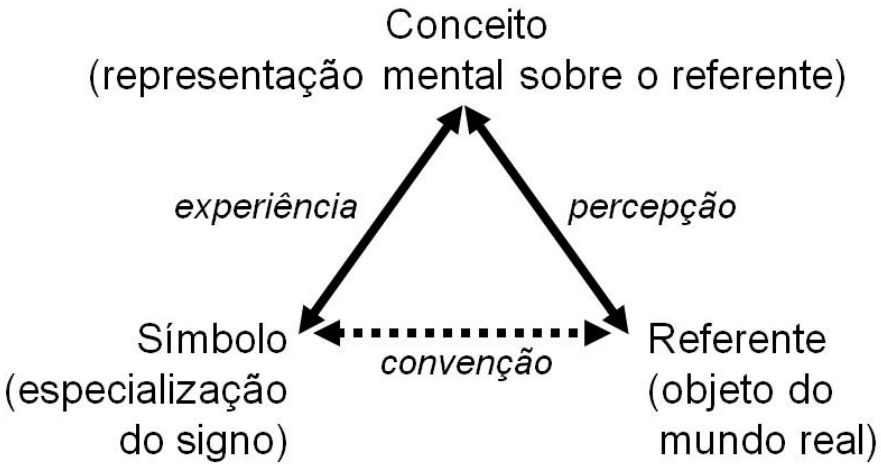


Figura 1: Triângulo do significado

Um termo necessariamente está associado a um conceito em uma linguagem especializada e pode ser constituído por uma ou mais palavras. Se o termo é formado por apenas uma palavra ele é denominado "termo simples", senão ele é denominado "termo complexo". Segundo o que prega a Teoria Geral da Terminologia (TGT), tanto o termo simples quanto o complexo estão associados a um conceito de maneira unívoca no domínio da linguagem especializada (Almeida, 2010). Independentemente da unidade lexical vir a se eleger a termo, para mineração de texto, o número de palavras ("n") que compõe uma unidade lexical mede sua gramatura, onde o termo simples, ou unidade lexical simples, possui o seu "n" de gramatura igual a um e é chamado de unigrama (nGram=1), se possuir duas palavras é chamado de bigrama (nGram=2), se possuir três palavras é chamado de trigrama e, assim sucessivamente

(Vieira; Lopes, 2010; Zerbinatti, 2010). Os termos complexos também podem representar agrupamentos linguísticos no idioma português como sintagmas nominais, se desempenham função de sujeito, sintagmas verbais, se desempenham função de predicado e, sintagmas preposicionais, se formados por preposição junto com sintagma nominal (Padilha; Lacerda, 2012).

A relação paritária (1:1) entre termo e conceito, proposto pela TGT e, ilustrado na figura 2, tem provocado muitas criticas por reduzir o estudo terminológico apenas ao estudo do termo o considerando como um rótulo e deixando de lado o texto completo (Pinto, 2010; Villavicêncio; Ramisch, 2010). Por consequência, a evolução da língua especializada e sua variação são negligenciadas (Silva-Filho, 2010). Para a Teoria Comunicativa da Terminologia (TCT) um conceito pode ser usado por vários domínios de conhecimento e, esta teoria se preocupa com o estudo da evolução dos conceitos porque os considera dinâmicos (Francelin; Kobashi, 2011).

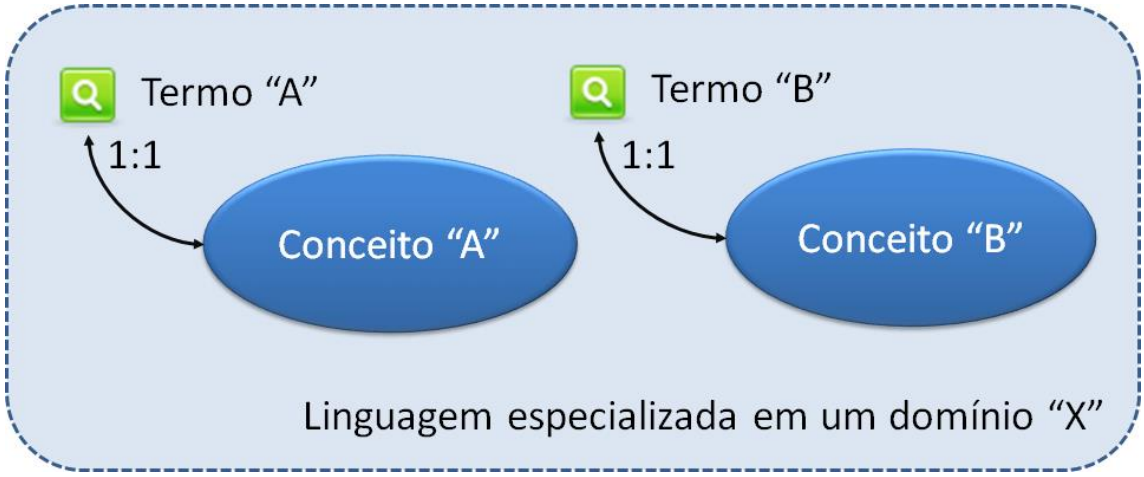


Figura 2: Para Teoria Geral da Terminologia um termo está associado a um único conceito.

1.1.3 Reconhecendo o termo

O estudo terminológico geralmente não é realizado por um membro da comunidade especializada, em geral, é o terminológo que realiza esta tarefa. Por não fazer parte da comunidade especializada ele não conhece em profundidade a linguagem especializada utilizada por esta comunidade dificultando a distinção entre uma unidade lexical comum de um termo. Um exemplo simples, que demonstra a dificuldade da semântica, é o termo “árvore” que no domínio de informática está associado ao conceito de “estrutura de dados hierárquica e complexa” e no domínio de botânica está associado ao conceito de “planta de grande porte” (Villavicêncio; Ramisch, 2010).

Não existem critérios absolutos para determinar quando uma unidade lexical é um termo ou quando não é dentro de um domínio especializado porque o termo utilizado na linguagem especializada não se difere da unidade lexical da língua geral. Todavia, existem algumas características que podem denunciar essas unidades lexicais especializadas (Almeida, 2010):

- Característica morfológica: quando o termo é marcado por um morfema greco-latino, exemplo: “osteófito”.

- Característica léxica: detectar derivações léxicas, exemplos: “osteófito”, “osteofitária” e “osteofitário”.

Embora as caracterizações morfológicas e lexicais ajudem a diferenciar o termo de uma unidade lexical, elas não são suficientes para identificar todos os termos de uma linguagem especializada porque nem todos possuem marcas formais, exemplo: “borda”, “cisto” e “nódulo”. Deste modo, é fundamental que faça parte da equipe do estudo terminológico um especialista da atividade específica que está sendo estudada.

1.1.4 Superestrutura e partes constituintes (categorias textuais)

Quando um texto possui uma estrutura esquemática global que o caracteriza, diz-se, que ele possui uma superestrutura textual. A superestrutura textual é composta por áreas de texto bem definidas chamadas categorias textuais ou partes constituintes do texto (Vieira, 2010).

Um texto constituído pelas partes introdução, objetivo, metodologia, resultado, discussão, conclusão e referências bibliográficas; um pesquisador o reconhece como um artigo científico porque a superestrutura textual o identifica como tal. Da mesma forma, quando um médico recebe um documento contendo os dados demográficos do paciente, médico solicitante, técnica do exame, descrição e impressão diagnóstica e o médico executante; ele o reconhece como um laudo de exame (figura 3).

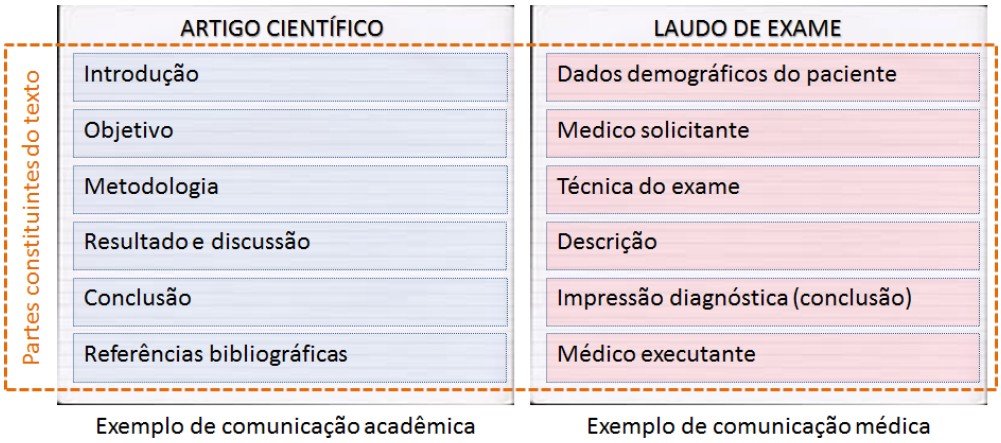


Figura 3: Exemplo de superestruturas textuais em diferentes atividades especializadas

O conhecimento da superestrutural textual facilita a generalização, a recordação e a reprodução do conteúdo além de definir a ordem e as relações entre suas partes constituintes (Vieira, 2010). Portanto, a própria superestrutura textual é uma informação relevante na comunicação especializada.

1.1.5 Atribuindo semântica ao pulso elétrico (ciclo)

A semântica é uma ciência que trata do estudo do significado e possui diversas abordagens geralmente complementares e não excludentes, mas é consenso que trata da significação (Almeida; Souza, 2011).

No inicio da computação era visível que o computador interpretava apenas pulsos elétricos porque existiam milhares de válvulas e interruptores para que um pequeno cálculo pudesse ser executado. Ao longo dos anos, diversas camadas de hardware e software foram acumuladas sobre o pulso elétrico para acessarem e controlarem os bits de, tal modo, que quando interagimos com o computador ele nos transmite a impressão de inteligência (vide figura 4).

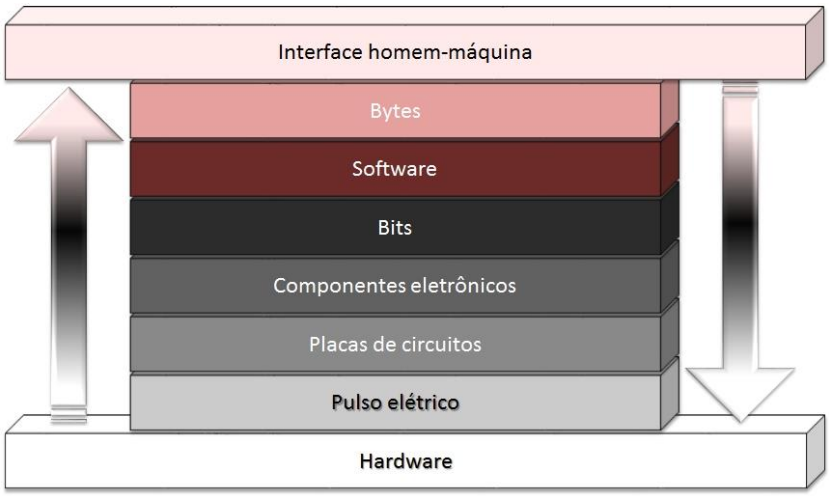


Figura 4: Impressão de inteligência: camadas sobre o pulso elétrico

Qualquer dado gravado seja um texto, uma imagem, um vídeo ou uma música, é constituído por um conjunto de símbolos que serão convertidos em bits e serão armazenados em um meio físico ocupando uma quantidade de bytes em um disco rígido, memória “flash” (“pen-drive”) ou sítios na internet, a rede mundial de computadores. Em todas as hipóteses, o dado será convertido em bits e armazenado

em bytes no meio eletrônico e, por este motivo, são necessários softwares que recuperem os bytes de dados armazenados, seja qual for o meio eletrônico.

1.1.6 Dados, informação e conhecimento.

O dado é um conjunto de símbolos, uma representação grosseria do mundo real, sem referencial que o explique (Cavalcante; Silva; Ferreira, 2011). Analise a data seguinte: “28/12/1976”. É possível afirmar que é uma data pretérita e está no formato brasileiro (DD/MM/AAAA), mas o que ela representa no mundo real? Seria uma data de nascimento? A data de um registro? A data de um fato histórico? A data de um atendimento clínico? Se não se conhece o que esta data representa no mundo real, então, ela é um dado.

A informação é um dado com significado, com referência no mundo real. Quando se sabe o que o dado referencia no mundo real, então, temos uma informação. Por exemplo, se afirmarmos que a data “28/12/1976” indica o nascimento de um ser humano, então, ela deixa de ser um dado e é promovida a instância da informação data de nascimento.

O conhecimento é o ato ou efeito de conceitualizar. A conceitualização é uma forma de representar os objetos, conceitos e outras entidades que existem em um domínio de interesse (contexto) e os relacionamentos semânticos entre estes objetos, conceitos e entidades (Nardon, 2003; Schiessl & Bräscher, 2011). Para chegar ao conhecimento são necessárias duas etapas e, a inteligência, três etapas distintas: 1) o dado forma a base da informação; 2) através do raciocínio humano apoiado por tecnologias gera-se o conhecimento; 3) a inteligência está em aplicar racionalmente o conhecimento gerado (Freire *et al.*, 2012).

Apenas os dados podem ser armazenados e processados em suporte eletrônico, porque a informação, o conhecimento e a inteligência exigem mediação humana (Freire *et al.*, 2012). A impressão de inteligência é atribuída ao computador através da utilização de softwares (algoritmos) que utilizam estruturas de dados sofisticadas simulando comportamentos humanos como o raciocínio.

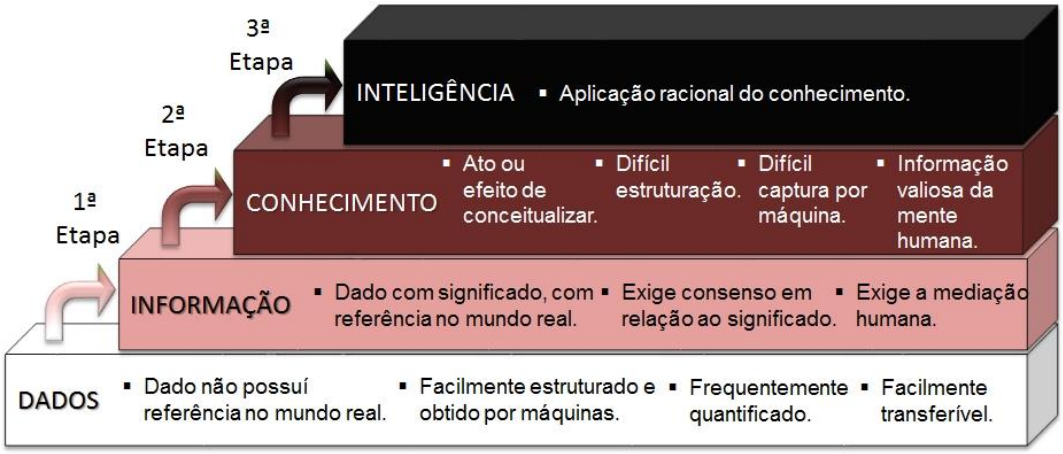


Figura 5: Etapas para geração do conhecimento e inteligência (adaptado de Freire, 2012)

A semântica computacional, a formalização eletrônica da informação, consiste em indicar ao computador através de estrutura de dados apropriada qual a informação que aquele dado representa no mundo real em um determinado contexto possibilitando ao ser humano recuperá-la.

Figura 6: Macro funcionamento de um sistema de recuperação de informação

1.1.7 Anonimização

O dado que permite identificar um individuo ou o torná-lo identificável é chamado de dado pessoal. É um preceito médico o sigilo sobre os dados pessoais do paciente e por este motivo nenhum dado deste tipo pode ser exposto publicamente porque além de propiciar a discriminação do paciente pode proporcionar diversos tipos de modelo de controle social e econômico. Existem cinco categorias de dados pessoais: 1) Dados anônimos, obtidos sem qualquer ligação com sua origem, ou seja, não identifica o indivíduo; 2) Dados anonimizados, são aqueles que quando foram obtidos permitiam a identificação do individuo e, cuja ligação, com este foi destruída impossibilitando sua identificação; 3) Dados dissociados, são dados que podem identificar o individuo com esforço considerável para consegui-lo; 4) Dados relativos a um individuo identificável; 5) Dados relacionados diretamente a um individuo (Schaefer, 2012).

Os dados pessoais que podem vir a identificar ou expor o paciente ou os profissionais envolvidos no estudo devem ser processados e anonimizados para sustentar os pilares da segurança da informação e do sigilo profissional médico. As técnicas computacionais que garantem a segurança da informação utilizam algoritmos de criptografia (Quantin *et al.*, 2009).

1.2 Mineração de texto

Por definição, Mineração de Texto (MT) é um conjunto de técnicas e processos para descoberta de novos conhecimentos em textos e, é uma subárea da Inteligência Artificial (Conrado; Gutiérrez; Rezende, 2012). Por se tratar de um conjunto de técnicas e processos e possuir um propósito específico a MT é considerada um método capaz, por exemplo, de extrair termos candidatos a conceitos e categorizar texto de documentos em suporte eletrônico (Rahman, 2011).

Este método surgiu na década de 80, devido ao aumento exponencial de documentos (informações) em suporte eletrônico e vem crescendo devido ao imenso avançado de tecnologia de hardware e software, que disponibiliza diferentes tipos de dados em suporte eletrônico. O seu objetivo é processar os dados semi-estruturados e não estruturados (texto livre) apontando o “conhecimento”[2] nele contido. Existem diversas abordagens de análise textual que fazem uso da MT e utilizam suas técnicas dependendo do propósito almejado pelo estudo (Elder; Miner; Nisbert, 2012), como ilustra a figura 7. Frequentemente as abordagens têm por propósito apontar informações ocultas em texto não estruturado e, por conta disso, a MT também é

[2] Como é discutido por este trabalho o conhecimento não está nas unidades lexicais, nos termos candidatos, porque eles apenas referenciam no ser humano um conhecimento tacito e, portanto, subjetivo. Para explicitar o conhecimento de forma computável é necessária a utilização de ontologia.

conhecida como descoberta de conhecimento em textos (em inglês: *Knonwledge Discovery in Text – KDT*) (Aggarwal; Cheng Xiang Zhai, 2012; Passini, 2012).

Figura 7: Abordagens e técnicas que fazem uso da mineração de texto (adaptado de Miner & et.al., 2012)

O processo de minerar texto em domínios de conhecimento especializados pode ser sintetizado em quatro etapas geralmente utilizadas (vide figura 8): 1) coleta de textos, o conjunto de documentos (textos) em suporte eletrônico de um determinado domínio é chamado corpus textual (Bubenhofer & et.al., 2012) e, é com base no corpus textual que o processo de MT se inicia; 2) pré-processamento, padroniza a tabela eletrônica de caracteres a ser utilizada, por exemplo, UniCode (Kahn, 2009), gera as unidades lexicais não normalizadas as armazenando em uma estrutura de dados conhecida como índice invertido ou arquivo invertido; 3) extração de padrões, utilizam-se algoritmos para extrair padrões e gerar os termos candidatos; 4) análise e avaliação dos resultados, os termos candidatos são avaliados de acordo com critérios estabelecidos pelo estudo, por exemplo, por frequência que apareceram

no corpus; 4) validação, através de análise subjetiva dos termos ou inferências de precisão (*precision*) e revocação (*recall*) (Duque, 2012; Oliveira; Motta; Cunha, 2011).

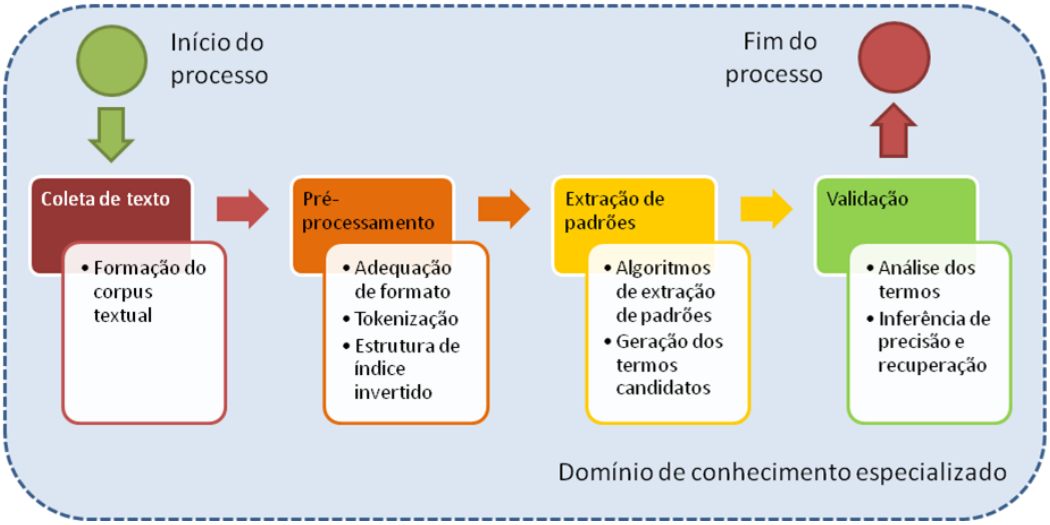


Figura 8: Processo de mineração de texto frequentemente utilizado em domínio de conhecimento especializado

Existem variações na definição do processo de MT porque cada trabalho faz adaptações de acordo com o propósito do estudo (Duque; Matos; Ciferri, 2012; Mhamdi; Elloumi, 2008; Oliveira; Motta; Cunha, 2011; Padilha; Lacerda, 2012; Passini, 2012; Rezende; Marcacini; Moura, 2011). Neste trabalho, por exemplo, a etapa de validação não será por inferência porque o propósito deste estudo não é a recuperação de documentos ou de informação, mas sim, a construção de uma ontologia para comparação de aspectos informacionais e terminológicos entre diferentes instituições. Outra adaptação é o processo de MT que será cíclico, permitindo o refinamento do produto de MT até que ele esteja satisfatório, conforme mostra a figura 9. Processos de mineração de texto em ciclo foram identificados por outros pesquisadores como Mhamdi e Elloumi em 2008, Rezende em 2011 e Duque em 2012 (Duque; Matos; Ciferri, 2012; Mhamdi; Elloumi, 2008; Rezende; Marcacini; Moura, 2011).

A mineração de texto é essencial para este trabalho porque propiciará a extração dos termos candidatos do corpus textual constituído por laudos radiológicos de ressonância magnética da coluna lombossacra em suporte eletrônico.

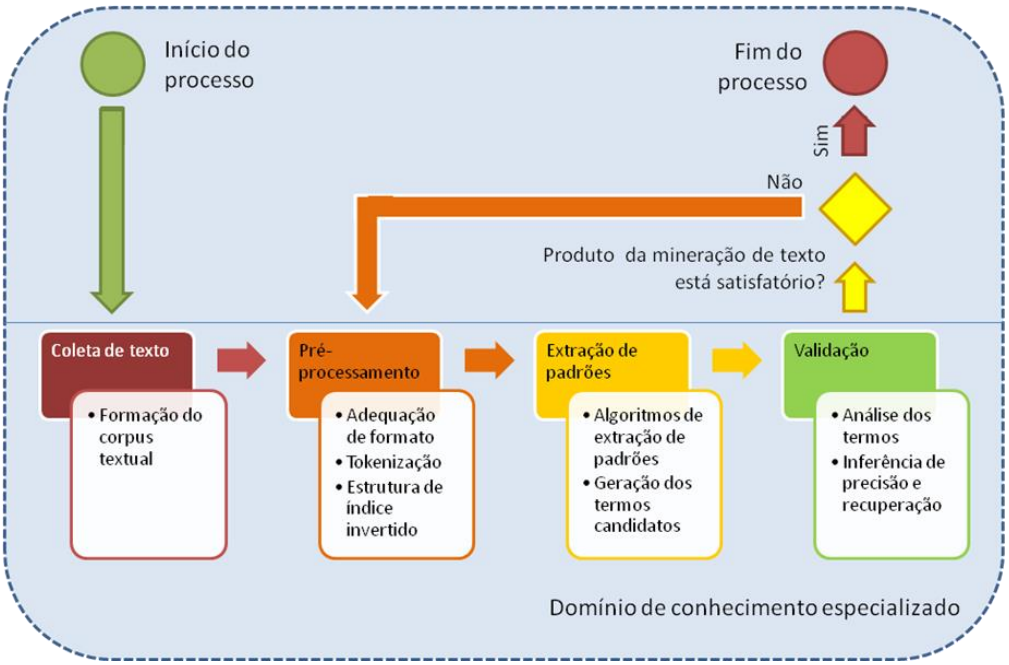


Figura 9: Processo de mineração de texto em ciclo frequentemente utilizado em domínio de conhecimento especializado

1.2.1 Decomposição do texto em parágrafos, frases e unidades lexicais

Um texto é composto de letras, que formam as palavras, números e símbolos agrupados em frases e parágrafos. Decompor esses elementos, significa percorrer o texto coletando a cada espaço a unidade lexical imediatamente posterior, a cada ponto final (“.”) a frase e a cada quebra de linha, oculta no texto, o parágrafo.

Geralmente a unidade lexical bruta coletada é uma palavra, mas pode ocorrer de ser uma preposição, um artigo, um sinal de pontuação ou mesmo um símbolo, por este motivo, informatas comumente empregam o termo “token” (do inglês, símbolo ou sinal) para unidade lexical bruta e “tokenização” para o processo de coletá-las. O separador, delimitador, comumente usado é o espaço ou sequencia de espaços ou

tabulação, todavia, é possível existir o comprimento arbitrário definido por “marcadores” no texto, como asterisco (“*”), pipeline (“|”), etc.. (Hussain; Afzal; Waqas, 2012).

Analisando os marcadores de formatação contidos em um fragmento de texto de laudo radiológico da coluna lombossacra, ilustrado na figura 10, fica evidente que os espaços são os limitadores comuns entre as unidades lexicais, o ponto final sinaliza o fim da frase e, ao final de cada parágrafo, existe um marcador de quebra de linha. Considerando como regra os marcadores de formatação, é possível executar o processamento de linguagem natural (PLN) e identificar os pedaços do texto correspondentes às unidades lexicais, frases e parágrafos.

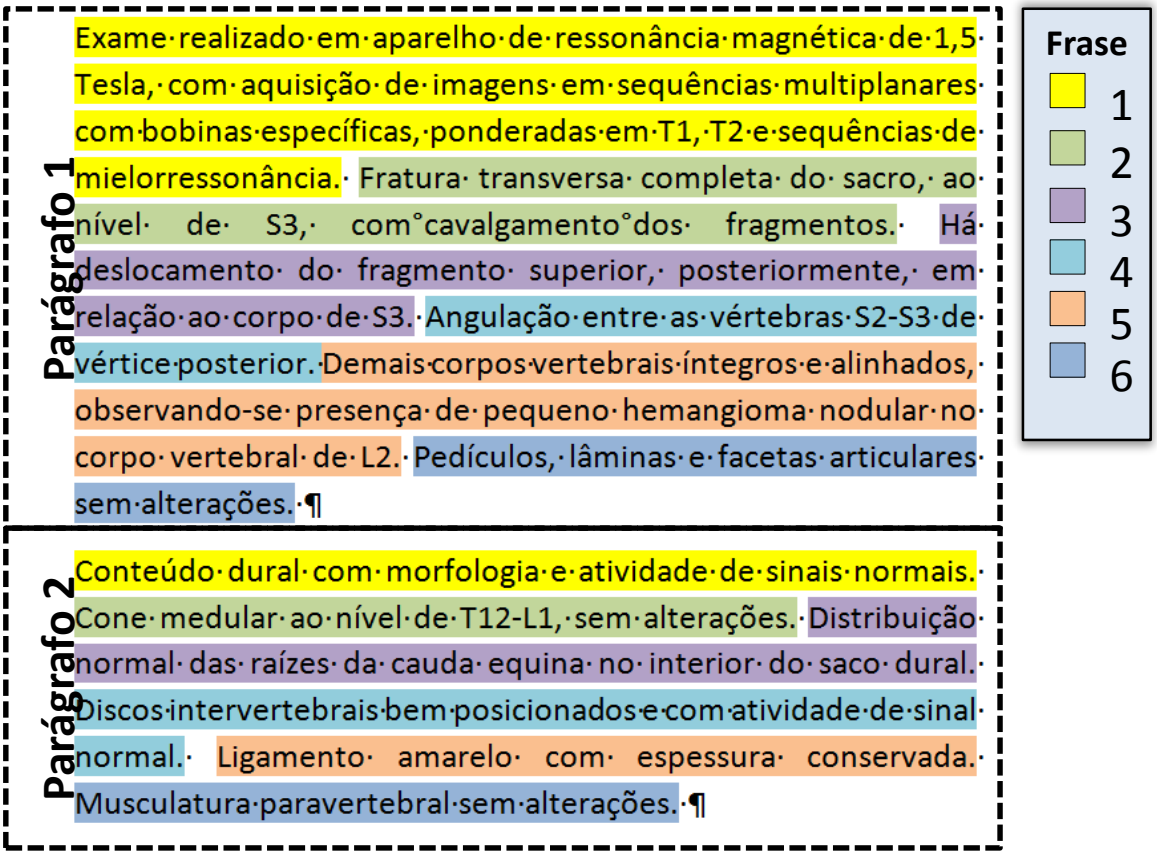


Figura 10: Fragmento de um laudo radiológico da coluna lombossacra exibindo marcadores de formatação

No exemplo ilustrado na figura 10, o PLN geraria 133 unidades lexicais brutas, distribuídas em doze frases e dois parágrafos. Conhecendo a posição de cada unidade lexical bruta dentro da frase e do parágrafo é possível determinar onde fica sua posição exata no texto permitindo estabelecer se a ocorrência foi no início, no meio ou no fim do texto. Por exemplo, a unidade lexical “1,5” está na nona posição na primeira frase do primeiro parágrafo; enquanto a unidade lexical bruta “T12-L1,” está na centésima posição na segunda frase do segundo parágrafo.

O mapeamento das unidades lexicais contidas nos documentos de um corpus textual fica armazenado em uma estruturada de dados chamado arquivo invertido ou índice invertido. Esta estrutura é capaz de rastrear os documentos onde determinada unidade lexical foi encontrada.

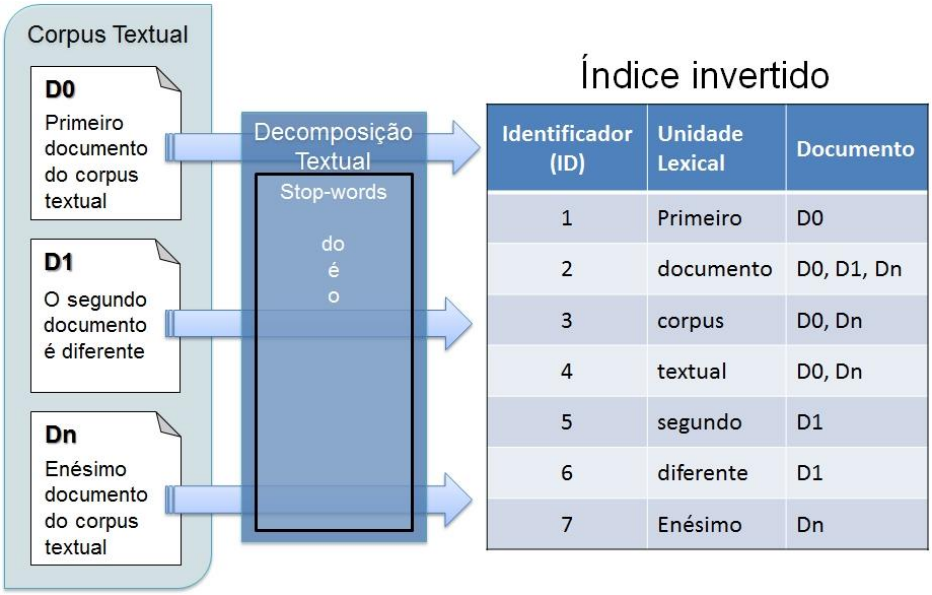


Figura 11: Modelo conceitual para geração de um índice invertido

Embora o modelo conceitual do índice invertido seja apresentado como uma matriz, esta estrutura de dados possui varias dimensões e, portanto, trata-se de uma matriz multidimensional. A coluna documento, por exemplo, precisa armazenar um número indefinido de documentos, além disto, existe a necessidade de armazenar

informações sintetizadas da frequência que unidade lexical ocorre independente do documento que foi encontrada e dentro do documento que foi encontrada. Desta forma, o modelo conceitual do índice invertido possui minimamente 2 dimensões: 1) Unidade lexical univocamente identificada; 2) Documentos de onde foram extraídos. Se as informações sintetizadas de frequência forem necessárias, então, duas dimensões devem ser adicionadas transformando o índice invertido em uma matriz de quatro dimensões: 1) Unidade lexical univocamente identificada; 2) Síntese da frequência das unidades lexicais; 3) Documentos de onde foram extraídos; 4) Síntese da frequência das unidades lexicais dentro do documento de onde foram extraídos.

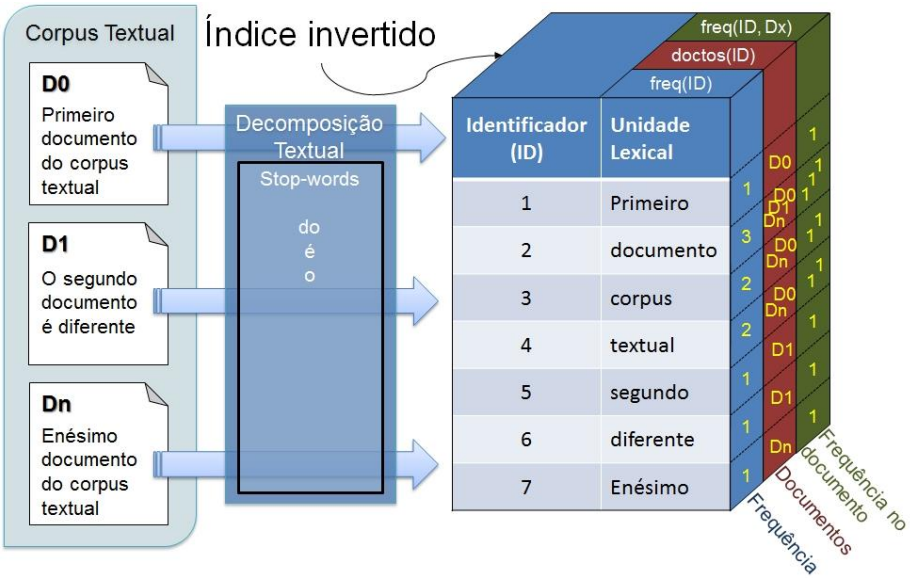


Figura 12: Modelo conceitual para geração de um índice invertido multidimensional

Acrescentando as dimensões de síntese da frequência é possível prever algumas funções como o retorno da frequência de uma determinada unidade lexical onde o identificador (ID) da unidade lexical que se deseja obter a frequência é passado e a frequência retornada (“freq(ID)”, exemplo “freq(2)=3”), a lista dos documentos de onde a unidade lexical foi extraída (“doctos(ID)”, exemplo “doctos(2)={‘D0’, ‘D1’,

‘Dn’}”), a frequência de determinada unidade lexical dentro de um determinado documento (“freq(ID, Dx)”, exemplo “freq(2, ‘D1’)=1”).

A Terminologia institui a “ficha terminológica” como o principal registro de dados terminológicos e é composta por um conjunto mínimo de informações, a saber: área temática que o conceito pertence, o(s) idioma(s), a definição do conceito, o mapeamento dos documentos de onde a unidade lexical foi extraída (provas textuais) e as fontes que documentam essa informação (Pavel; Nolet, 2002). Portanto, o índice invertido seria utilizado dentro de uma ficha terminológica eletrônica porque ele cumpre o papel de mapear a origem das unidades lexicais determinando quais as provas textuais (documentos) de onde eles foram extraídos.

1.2.2 Normalização léxica

O processo de decomposição do texto é aplicado sobre um corpus textual escrito em um determinado idioma e, portanto, o resultado são unidades lexicais brutas que obedeceram as regras gramaticais do idioma onde foram escritas e, por conta disso, carregam consigo as variações morfológicas do léxico (Vasilévski, 2011).

Normalização léxica é uma importante tarefa do processo de mineração de texto porque seu objetivo é minimizar a variação morfológica, ou seja, variação léxica, do conjunto de léxico e provocar aglutinação das unidades lexicais, normalização do léxico. Esta tarefa é executada após ou junto com a decomposição do texto e dependendo da técnica aplicada pode eliminar os símbolos especiais e de pontuação da linguagem que ficaram concatenados as unidades lexicais brutas, convertê-las para minúsculo, remover acentos, prefixos e sufixos, converter formas nominais para verbos, etc. (Braga, 2011).

Existe uma diferença importante entre variação léxica e variação terminológica. A variação léxica diz respeito unicamente às diversas formas, variação morfológica,

que um léxico pode assumir por conta das regras linguísticas do idioma que está submetido, por exemplo, um mesmo léxico será escrito de formas diferentes no singular, plural, gerúndio, tempo verbal, etc. A variação terminológica, por outro lado, se refere à problemática do conceito associado ao termo observando fenômenos como a polissemia, onde um termo possui vários significados, e a sinonímia, onde vários termos são atribuídos ao mesmo significado (Corrêa; Miranda, 2011; Faulstich, 1995). A normalização reduz a variação léxica, mas não reduz a variação terminológica.

A fim de exemplificar a tarefa de normalização léxica e a criação da lista de termos candidatos foram aplicadas as técnicas de conversão para minúsculo, remoção de acento e caracteres de pontuação e conversão para singular no fragmento de laudo radiológico “A” (o resultado deste ensaio pode ser verificado na tabela 1).

	Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências
1	de	12	23	conservada	1	45	cauda	1	67	ligamento	1
2	com	6	24	especifica	1	46	cavalgamento	1	68	magnetica	1
3	e	5	25	l2	1	47	completa	1	69	medular	1
4	do	4	26	mielorressonancia	1	48	cone	1	70	morfologia	1
5	em	4	27	pediculo	1	49	conteudo	1	71	multiplanar	1
6	alteracao	3	28	posicionado	1	50	demais	1	72	musculatura	1
7	normal	3	29	posterior	1	51	deslocamento	1	73	nodular	1
8	ao	3	30	posteriormente	1	52	disco	1	74	observando-se	1
9	corpo	3	31	raiz	1	53	distribuicao	1	75	paravertebral	1
10	sem	3	32	sacro	1	54	entre	1	76	pequeno	1
11	da	2	33	superior	1	55	equina	1	77	ponderada	1
12	dural	2	34	t1	1	56	espessura	1	78	presença	1
13	fragmento	2	35	t12-l1	1	57	exame	1	79	realizado	1
14	s3	2	36	tesla	1	58	faceta	1	80	relacao	1
15	sequencia	2	37	amarelo	1	59	fratura	1	81	ressonancia	1
16	atividade	2	38	angulacao	1	60	ha	1	82	s2-s3	1
17	nivel	2	39	aparelho	1	61	hemangioma	1	83	saco	1
18	no	2	40	aquisicao	1	62	imagens	1	84	t2	1
19	sinal	2	41	articular	1	63	integro	1	85	transversa	1
20	vertebral	2	42	as	1	64	interior	1	86	vertebra	1
21	1,5	1	43	bem	1	65	intervertebral	1	87	vertice	1
22	alinhado	1	44	bobina	1	66	lamina	1			

Tabela 1: Lista dos termos candidatos extraídos do fragmento de laudo radiológico "A"

Analisando os termos candidatos extraídos do fragmento de laudo radiológico “A” é possível constatar que houve uma redução da quantidade de unidades lexicais

de 133 para 87 (vide tabela 1). Isto ocorreu porque as unidades lexicais brutas tiveram o seu léxico normalizado pelas técnicas descritas anteriormente provocando a fusão, a aglutinação, das unidades lexicais normalizadas (termos candidatos) e permitindo a contagem de repetições (ocorrência) de cada uma delas. Antes da normalização a unidade lexical “normal” e “normais” seriam diferentes, com a conversão para singular “normais” foi convertida para “normal” e por isto a quantidade de ocorrências apurada foi três ao invés de dois.

Com o objetivo de realizar inferências sobre corpus textuais, Roberto & et.al. em 2012, reuniram diversas medidas de baixo custo computacional, independentes do idioma e domínio do texto, que permite avaliar o quanto um texto é fecundo, o quanto ele é rico, ou seja, avaliar sua riqueza léxica (Roberto; Martí; Llorente, 2012). Entre os indicadores que avaliam a riqueza léxica está a densidade léxica, que se traduz pela razão entre a quantidade de unidades lexicais normalizadas (termos candidatos) (*Nlex*) e a quantidade de unidades lexicais brutas (*N*) de um corpus textual (vide equação 1) indicando a proporção de léxicos distintos existente no texto.

$$DL = \frac{NLex}{N}$$

Equação 1: Definição do índice de densidade léxica (DL)

A razão entre a diferença da quantidade de unidades lexicais brutas (*N)* e a quantidade de termos candidatos (*Nlex*) pela quantidade de unidades lexicais brutas (*N*) nos permite calcular a aglutinação léxica que indica o quanto de variação léxica existe (vide equação 2). A aglutinação léxica é o valor complementar à densidade léxica e, portanto, inversamente proporcional à mesma.

$$AL = (1-DL)$$

Equação 2: Definição do índice de aglutinação léxica (AL)

Aplicando o calculo de densidade léxica e aglutinação léxica no exemplo de fragmento de laudo radiológico “A” a variável *Nlex* assumiria a quantidade de unidades lexicais normalizadas (termos candidatos), do exemplo 87 e, a variável *N* assumiria a quantidade de unidades lexicais brutas, do exemplo 133, portanto a densidade léxica (DL) do exemplo é igual a DL=87÷133= 0,6541 e a aglutinação léxica (AL) é igual a AL=(133-87) ÷133=0,3458 (vide tabela 2).

Análise da Riqueza Léxica	
Quantidade de unidades lexicais brutas	133
Quantidade de termos candidatos	87
Densidade Léxica (DL)	0,6541
% Densidade Léxica	65%
Aglutinação Léxica (AL)	0,3458
% Aglutinação	35%

Tabela 2: Análise da riqueza léxica do fragmento de laudo radiológico "A"

Analisando a riqueza léxica do fragmento de laudo radiológico “A” é possível afirmar que 35% das unidades lexicais são constituídas por algum tipo de variação léxica e que os 65% restantes são unidades lexicais normalizadas distintas. Portanto, as unidades lexicais brutas geradas pela decomposição do texto darão origem aos termos candidatos após sua normalização e, quando estes forem associados aos seus respectivos conceitos serão eleitos “termos” da linguagem especializada.

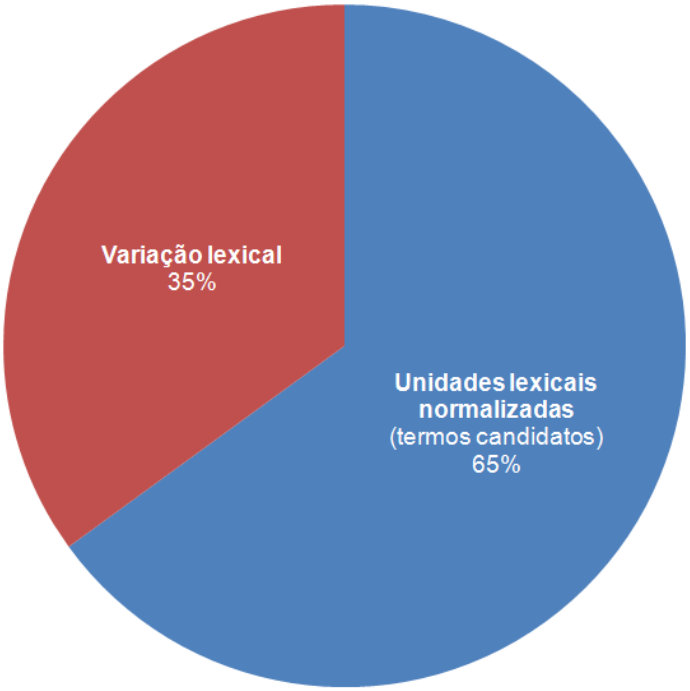


Gráfico 1: Análise da riqueza léxica do fragmento de laudo radiológico "A"

1.2.3 Normalização lexical e a lista de Stopwords

Observando a tabela 1, é possível constatar que os termos candidatos com maior ocorrência extraídos do fragmento de laudo radiológico "A" são unidades lexicais utilizadas para unir as palavras na oração, ou seja, são preposições: "de", "com", "e", "do" e "em". As preposições não possuem referência, significado, no mundo real e por este motivo compõe a lista de unidades lexicais (palavras) a serem desprezadas pela normalização lexical denominada: "lista de *stopwords*". Além das preposições os artigos, pronomes e advérbios são bons exemplos de classes de palavras que não possuem referencia no mundo real e estão no texto apenas para dar sentido a frase e cumprir um papel sintático na língua que foram escritos (Rezende; Marcacini; Moura, 2011). Por conta da dependência do idioma existirá uma lista diferente de *stopwords* para cada língua.

O primeiro passo para utilização de lista de *stopwords* é a definição de quais serão as unidades lexicais que irão compor esta lista. Com base na tabela 1, é possível estabelecer uma lista de *stopwords* composta por 11 palavras, conforme mostrado na tabela 3.

	Stopwords		Stopwords		Stopwords
1	de	5	em	9	as
2	com	6	ao	10	demais
3	e	7	da	11	entre
4	do	8	no		

Tabela 3: Lista de stopwords extraída do fragmento de laudo radiológico "A"

Para exemplificar a utilização de lista de *stopwords* elas serão removidas da lista de termos candidatos extraídos do fragmento de laudo radiológico “A” e uma nova lista de termos candidatos e analise da riqueza léxica será confeccionada.

	Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências		Termos Candidatos	Contagem de Ocorrências
1	alteracao	3	23	raiz	1	45	espessura	1	67	presença	1
2	normal	3	24	sacro	1	46	exame	1	68	realizado	1
3	corpo	3	25	superior	1	47	faceta	1	69	relacao	1
4	sem	3	26	t1	1	48	fratura	1	70	ressonancia	1
5	dural	2	27	t12-l1	1	49	ha	1	71	s2-s3	1
6	fragmento	2	28	tesla	1	50	hemangioma	1	72	saco	1
7	s3	2	29	amarelo	1	51	imagens	1	73	t2	1
8	sequencia	2	30	angulacao	1	52	integro	1	74	transversa	1
9	atividade	2	31	aparelho	1	53	interior	1	75	vertebra	1
10	nivel	2	32	aquisicao	1	54	intervertebral	1	76	vertice	1
11	sinal	2	33	articular	1	55	lamina	1			
12	vertebral	2	34	bem	1	56	ligamento	1			
13	1,5	1	35	bobina	1	57	magnetica	1			
14	alinhado	1	36	cauda	1	58	medular	1			
15	conservada	1	37	cavalgamento	1	59	morfologia	1			
16	especifica	1	38	completa	1	60	multiplanar	1			
17	l2	1	39	cone	1	61	musculatura	1			
18	mielorressonancia	1	40	conteudo	1	62	nodular	1			
19	pediculo	1	41	deslocamento	1	63	observando-se	1			
20	posicionado	1	42	disco	1	64	paravertebral	1			
21	posterior	1	43	distribuicao	1	65	pequeno	1			
22	posteriormente	1	44	equina	1	66	ponderada	1			

Tabela 4: Lista dos termos candidatos extraídos do fragmento de laudo radiológico "A" considerando a lista de "stopwords"

A tabela 4 mostra a lista de termos candidatos após a remoção das unidades lexicais existentes na lista de *stopwords* do fragmento de laudo radiológico “A”. As alterações nos indicadores de densidade e aglutinação léxica podem ser verificados na tabela 5.

Análise da Riqueza Léxica	
Quantidade de unidades lexicais brutas	133
Quantidade de termos candidatos	76
Densidade Léxica (DL)	0,5714
Aglutinação Léxica (AL)	0,4286
% Densidade Léxica	57%
% Aglutinação	43%

Tabela 5: Análise da riqueza léxica do fragmento de laudo radiológico "A" considerando lista de "stopwords".

Ao considerar a lista de *stopwords* na análise da riqueza léxica do fragmento de laudo radiológico “A” o percentual de densidade léxica diminuiu de 65% para 57% consequentemente aumentando o percentual de aglutinação de 35% para 43% porque a quantidade de termos candidatos foi reduzida pela remoção das *stopwords* enquanto a quantidade de unidades lexicais brutas se manteve a mesma.

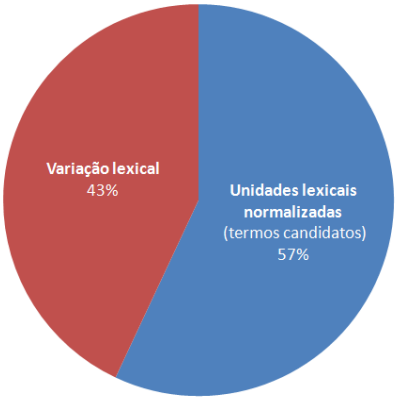


Gráfico 2: Análise da riqueza léxica do fragmento de laudo radiológico “A” considerando "stopwords"

1.3 Ontologia

A primeira etapa para geração do conhecimento e inteligência é transformar dados em informação. Mas não basta descrever o que uma unidade lexical representa no mundo real, pois para transformá-la em uma informação, em um “termo”, é preciso associá-la a um conceito e, este sim, deve descrever o objeto do mundo real. Isso é necessário porque o “termo”, unidade lexical que representa o conceito, é sensível ao idioma e aos fenômenos linguísticos, como a polissemia (uma palavra com muitos significados) e sinonímia (várias palavras atribuídas ao mesmo significado) (Corrêa; Miranda, 2011).

O conceito é um agrupamento de termos e pode estar associado a diversos deles inclusive de diferentes idiomas dentro de um mesmo contexto. Dentro de um produto terminológico, o conceito precisa ser tão unívoco quanto é o objeto que ele representa no mundo real em determinado contexto, pois se está condição não for atendida, certamente ocorrerão inconsistências informacionais. Além do objeto observado no mundo real, é possível conceituar a classe a qual o objeto pertence observando suas características, semelhantes e diferenças, por exemplo, “espondilose” e “osteoartrite” são objetos do mundo real e ambas são instancias da classe de “doenças degenerativas”. Produtos terminológicos reconhecidos mundialmente, como o SNOMED-CT, são orientados a conceitos e, portanto, descrevem univocamente os objetos do mundo real agrupando os termos que remetem a eles independente do idioma (Ceusters; Smith, 2010).

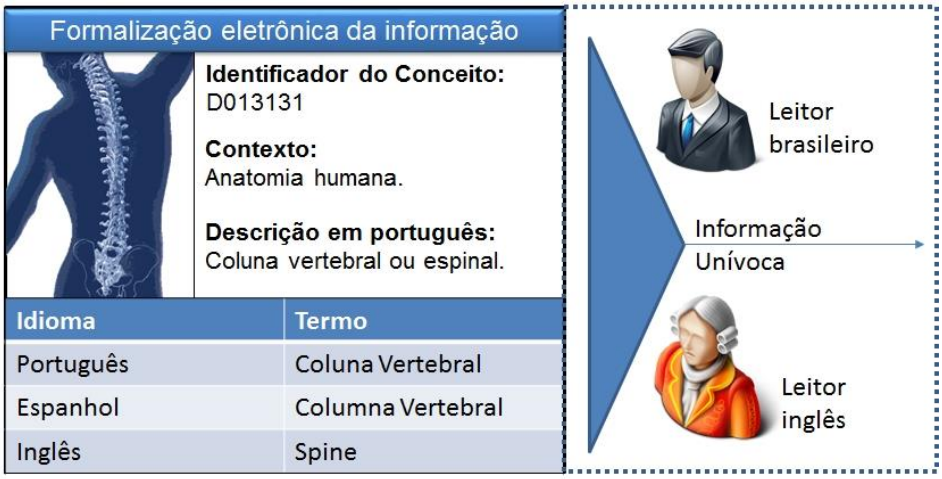


Figura 13: Formalização eletrônica do conceito: informação unívoca.

A formalização eletrônica da informação, ou seja, formalização do conceito possibilitará que dois leitores de idiomas diferentes consigam obter informação unívoca sobre o mesmo objeto do mundo real, mas não conseguirá atingir a segunda etapa da geração do conhecimento e da inteligência porque, para transformar informação em conhecimento, é necessário formalizar, explicitar, as relações existentes entre as informações, entre os conceitos. A técnica que permite formalizar conceitos e explicitar suas relações chama-se ontologia e a representação gráfica dessas relações se denomina diagrama ou árvore conceitual (Pavel; Nolet, 2002).

No inicio dos estudos filosóficos, havia uma vertente que se interessava pelo mundo intangível chamada de *metaphysica specialis* e, outra, que tratava de assuntos relacionados ao mundo tangível chamada *metaphysica generalis*. A primeira se ocupava com assuntos relacionados à existência de entidades como Deus, enquanto a segunda, se ocupava com a natureza e constituição da realidade, com sua estrutura e com os conceitos mais gerais do “ser”. Mais tarde, a primeira, *metaphysica specialis*, originou estudos como a teologia, metafísica, etc.. Da mesma forma, a segunda, *metaphysica generalis*, originou a ontologia (Krause, 2011). Portanto, o termo “ontologia” é oriundo da filosofia.

Na década de 90, o termo “ontologia” foi apropriado pela computação na área de inteligência artificial em projetos de organização de grandes bases de conhecimento (Silva-Filho, 2010). Para computação ontologia é uma especificação formal e explícita de uma conceitualização (Bellatreche; Ait-Ameur; Chakroun, 2010; Borst; Akkermans; Top, 1997; Cimiano; Reyle; Šarić, 2005; Mhiri; Despres, 2007). O termo “formal” significa que a conceitualização deve ser legível para computadores (Zerbinatti, 2010). A conceitualização é uma forma de representar os objetos, conceitos e outras entidades que existem em um domínio de interesse (contexto) e o seu relacionamento semântico entre estes objetos, conceitos e entidades (Nardon, 2003).

1.3.1 Decisão ontológica contextual vertical e horizontal

A ontologia permite a formalização eletrônica da informação e a explicitação de sua conceitualização. Todavia, quando a conceitualização ocorre a partir de um corpus textual o conceito poderá ser definido de acordo com a sua posição no texto, na frase, no parágrafo, na categoria textual, ou seja, dentro do seu contexto (Pavel; Nolet, 2002).

A decisão ontológica contextual horizontal permite a um termo candidato referenciar um conceito dependendo do conceito que o preceder ou suceder seja ele um objeto ou uma classe. Diz-se horizontal porque, projetando um plano cartesiano sobre o texto, a dependência contextual com o texto está no eixo “X”. Por exemplo: se antes do termo candidato “sinal” estiver o conceito de “aumento”, referenciado pelos termos em português “alto”, “hiper” ou “aumento”, o conceito referenciado pelo termo candidato será “hipersinal”; se antes do termo candidato “sinal” estiver o conceito de “discreto”, referenciado pelos termos em português “baixo”, “reduzido” ou “discreto”, o

conceito referenciado pelo termo candidato será “hipossinal”; se nenhuma dessas condições ocorrerem, então, o termo candidato referenciará o conceito “sinal normal”.

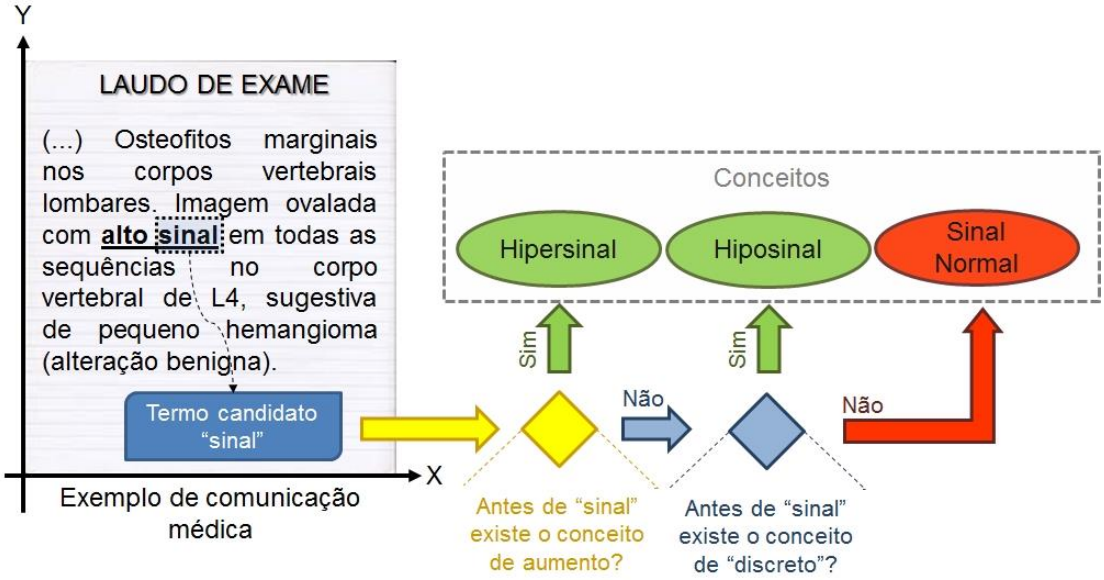


Figura 14: Exemplo de utilização da decisão ontológica contextual horizontal

A decisão ontológica contextual vertical permite a um termo candidato referenciar um conceito, seja ele um objeto ou uma classe, dependendo da sua posição vertical no texto. Na posição vertical se encontram as partes constituintes aderentes a superestrutura textual. Denominou-se decisão ontológica contextual vertical porque a dependência com o texto está no eixo “Y”. Por exemplo: se o termo candidato “aparelho” ocorrer no primeiro parágrafo do laudo radiológico em uma instituição que adote este parágrafo para descrever a técnica do exame (categoria textual implícita) este termo referenciará o conceito “equipamento”. Todavia, se este mesmo termo candidato ocorrer na categoria textual descrição ou impressão diagnóstica (conclusão) do laudo do exame o conceito referenciado será “conjunto de órgãos” porque estará se referindo ao conjunto de órgãos que compõe um determinado aparelho do corpo humano (por exemplo: “aparelho digestivo”, “aparelho auditivo”, etc.).

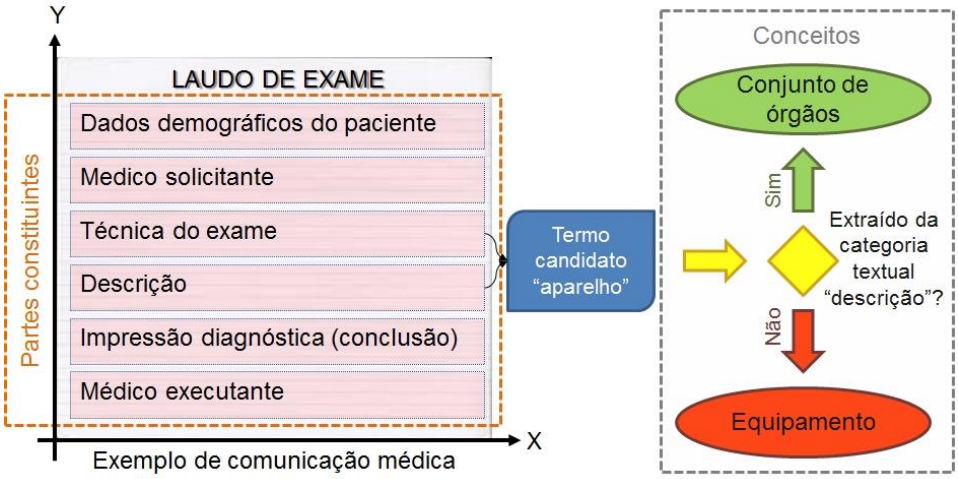


Figura 15: Exemplo de utilização da decisão ontológica contextual vertical

1.3.2 Breve análise das ferramentas computacionais existentes

Os seguintes requisitos foram considerados na analise das ferramentas computacionais: 1) minerar texto de corpus textual em português coletados em arquivo do Microsoft Word (.doc) e do Microsoft Excel (.xls); 2) construir ontologia apoiado pelo DeCS (Descritores de Ciências da Saúde) por se tratar de um produto terminológico em português no domínio da saúde; 3) suportar ontologia posicional; 4) utilizar banco de dados relacional (RDB – Relational DataBase) (Kong; Zhang, 2011); 5) Possibilitar anonimizar ou dissociar os dados que tornam as pessoas envolvidas no laudo identificáveis; 6) Visualização da ontologia em árvore.

Ferramenta	Referência	Contempla Português	Processa Corpus Textual	Usa RDB	Gera termos candidatos	Anonimiza dados pessoais	Conceitualiza	Arvore Ontologica	Integração com DecS
TextAlyser	Klemann, 2011	Não	Não	Não	Sim	Não	Não	Não	Não
WordCounter	Klemann, 2011	Não	Não	Não	Sim	Não	Não	Não	Não
TagCrowd	Klemann, 2011	Sim	Não	Não	Sim	Não	Não	Não	Não
Sobek	Klemann, 2011	Sim	Não	Não	Sim	Não	Não	Sim	Não
Pretext	Conrado & et.al., 2012	Sim	Sim	Não	Sim	Não	Não	Não	Não
Protegé	Deus & Cabral, 2011	Não	Não	Não	Não	Não	Sim	Sim	Não

Tabela 6: Ferramentas computacionais para mineração de texto e construção de ontologias.

A análise realizada das ferramentas computacionais para mineração de texto e construção de ontologias não é exaustivo, porém, mesmo desconsiderando o suporte a ontologia posicional, nenhuma destas ferramentas atendeu a todas as necessidades deste trabalho. Devido a este fator optou-se pela construção da ferramenta *meta-genaralis* descrita em anexo.

1.4 O exame de diagnóstico por imagem

Exames de imagens são métodos diagnósticos auxiliares na prática médica e são solicitados pelo médico que acompanha o paciente a fim de descartar ou comprovar sua hipótese diagnóstica. As imagens do paciente são adquiridas (geradas) por diversos equipamentos de alta densidade tecnológica como ressonância magnética, tomografia computadorizada, videoendoscopia, ultrassonografia, mamografia digital, raio-x digital, densitometria óssea, ecocardiografia, histeroscopia, etc. (Machado *et al.*, 2012). As imagens adquiridas por esses equipamentos são denominadas “imagens médicas”.

Diferente de um exame laboratorial, onde o médico que solicita é o mesmo que analisa e interpreta o resultado, o exame de diagnóstico por imagem exige um médico executante especialista em análise e interpretação das imagens médicas para

identificar o que elas revelam (Maciel-Lima; Rasia, 2012). O médico executante, chamado de radiologista ou imaginologista, é responsável por coordenar a execução do exame de diagnóstico por imagem e de registrar sua análise e interpretação em um relatório médico denominado “laudo radiológico”. Basicamente, o laudo radiológico deve conter: a técnica utilizada no exame, os dados do paciente, a região examinada, a descrição dos achados da imagem (Prestes-Júnior; Tourinho; Rangel, 2012) e, quando o diagnóstico é assertivo, uma conclusão ou impressão diagnóstica (Fenelon, 2003). Este relatório é enviado ao médico solicitante que concluirá ou continuará a investigação sobre o problema do paciente.

Quando o médico que cuida do paciente decide solicitar um exame de diagnóstico por imagem para fortalecer sua hipótese diagnóstica, ele inicia um processo que culmina na elaboração do laudo radiológico pelo médico executante e termina na verificação deste laudo pelo médico solicitante (figura 16).

A eficiência diagnóstica para os exames de imagem dependem da precisão de quem indica, da habilidade de quem realiza e o interpreta e da comunicação entre ambos (Pinto; Brunese, 2010). Todavia, habilidade em analisar imagem médica demanda competências específicas e depende da experiência e conhecimento adquirido pelo radiologista ao longo do tempo (Maciel-Lima; Rasia, 2012). Portanto, a atividade radiológica é uma atividade especializada e, como tal, possui uma linguagem especializada com termos associados a conceitos próprios do seu domínio de conhecimento.

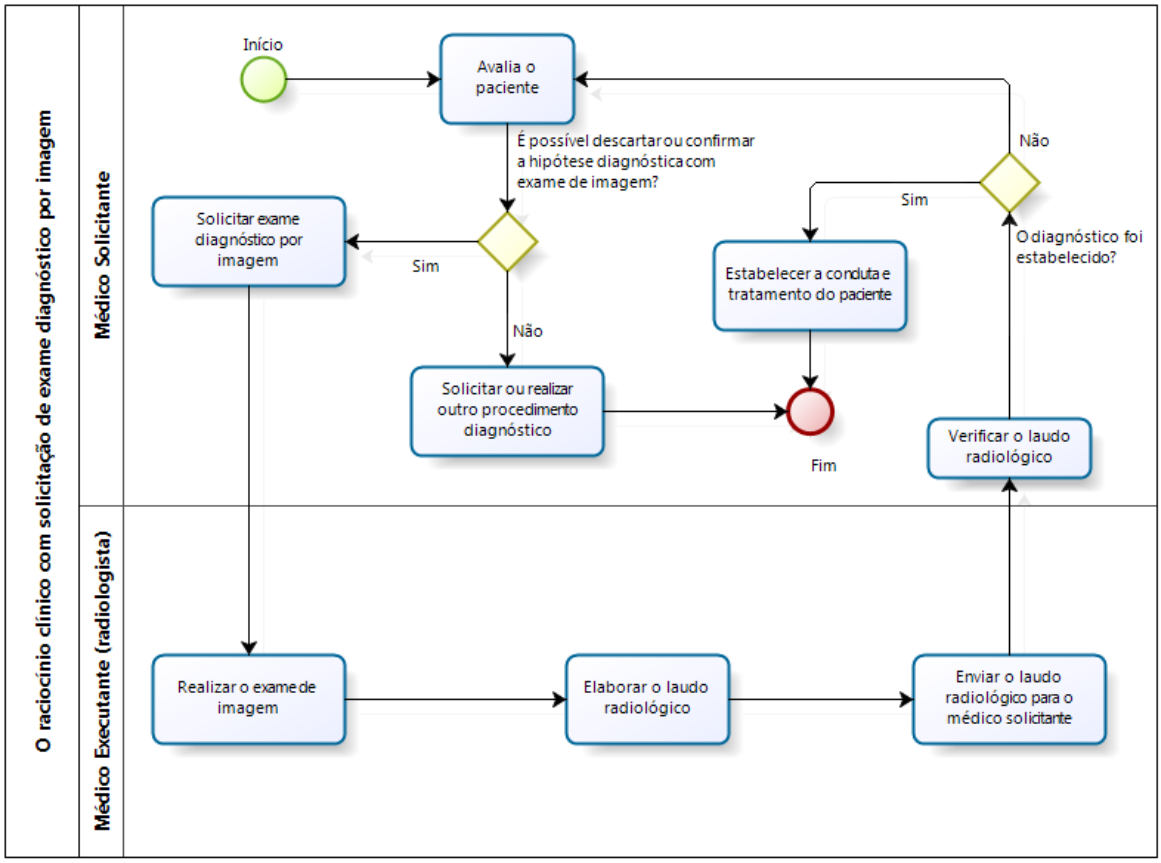


Figura 16: O raciocínio clínico com solicitação de exame diagnóstico por imagem

1.4.1 A Terminologia e o laudo radiológico

A característica marcante de uma atividade especializada é a sua necessidade de comunicação sem ambiguidade. A padronização terminológica diminui a variação das descrições dos achados no registro do laudo melhorando a transmissão das informações entre solicitante e executante (Barbosa *et al.*, 2010) consequentemente diminuindo erros decorrentes de interpretação inadequada.

O laudo radiológico estruturado propõe controle da terminologia empregada nas partes constituintes descrição e conclusão. A análise informacional destas partes constituintes pode levar a identificação e classificação de outras informações nelas contidas criando novas partes constituintes mais especificas e restritas a poucos termos. Estruturar um laudo radiológico significa “pinçar” de dentro das grandes partes constituintes a maior parte das informações especificas gerando partes constituintes menores com termos restritos àqueles que representam (categorizam) aquela informação. Estruturar um laudo radiológico não é trivial, porque as informações precisam atender o médico solicitante, que busca um diagnóstico conciso e preciso e, o médico executante que precisa descrever o que observou na imagem e embasou a sua impressão diagnóstica.

A estruturação dos laudos radiológicos, em geral, se realiza de forma empírica perguntando a cada uma das equipes, solicitantes e executantes, quais as partes constituintes são necessárias para compor o laudo radiológico. Desta forma, informações ou termos que são frequentemente empregados podem deixar de serem considerados, provocando rejeição da utilização do laudo estruturado. Não raro, médicos de outras regiões distintas do Brasil contestarem as informações de domínio pré-definidos (descritores) por estarem habituados com terminologia própria de sua região. O ideal seria existir uma metodologia apoiada por ferramenta computacional

capaz de destacar, dos laudos já confeccionados de forma livre ou semiestruturada, os termos candidatos mais frequentes, elencando-os a possíveis termos dentro de uma categoria textual, levando em consideração as especificidades da linguagem especializada em que o laudo está inserido, as especificidades da língua portuguesa e a cultura local.

1.4.2 Processos para elaboração de laudos radiológicos

Existem dois métodos para elaboração do laudo radiológico: por digitação ou transcrição ditada.

Por digitação ou método convencional, o radiologista escreve de próprio punho suas anotações (manuscrito) e entrega para um digitador transcrevê-lo em suporte eletrônico, no formato de um arquivo com texto possibilitando a impressão e assinatura do radiologista. Ainda é possível que o próprio radiologista digite o laudo, diretamente em suporte eletrônico e o imprima para assiná-lo dispensando o digitador. No Brasil, também há possibilidade do radiologista assinar digitalmente o laudo radiológico em suporte eletrônico, utilizando o seu certificado digital padrão ICP-Brasil que dispensa a assinatura em papel e por consequência a impressão do laudo.

Por transcrição ditada, o radiologista grava sua voz em fitas cassetes ou em suporte eletrônico utilizando arquivos digitais e entrega para um digitador ouvir e transcrever o ditado em suporte eletrônico. No início da década passada (2001), softwares comerciais capazes de interpretar a voz e transcrever automaticamente o texto foram construídos, porém, este tipo de software depende do vocabulário no domínio especializado e do idioma em que está inserido (Martins, 2011).

2 OBJETIVOS

2.1 Objetivo geral

Elaborar e avaliar uma metodologia para estruturação de informações de laudos radiológicos em texto livre que contemple as especificidades brasileiras, incluindo a terminologia empregada.

2.2 Objetivos específicos

Verificar a possibilidade de se explicitar estruturas informacionais e terminologia empregadas na confecção de laudos radiológicos em texto livre ou semiestruturado apoiado por produtos terminológicos eletrônicos;

Avaliar se as estruturas e terminologia explicitada e formalizada estão corretas consultando especialistas da área.

3 MATERIAL E MÉTODOS

3.1 Composição da amostra

Para confecção da metodologia, proposta, que poderá ser extrapolada para qualquer tipo de relatório médico e verificar o efeito de sua aplicação, selecionou-se uma amostra composta de 1701 (um mil e setecentos e um) laudos consecutivos de ressonância magnética de coluna lombossacra de três instituições: **A**) Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (HCFMRP-USP), **acadêmica** e no estado de São Paulo (**SP**); **B**) Clínica CEDIRP de Ribeirão Preto, **privada** e no estado de São Paulo (**SP**); **C**) Omnimagem de Fortaleza, **privada** e no estado do Ceará (**CE**).

Instituição	Tipo	Período de emissão dos laudos			Estado da União	Quantidade de laudos coletados
		Ano	Mês			
			Inicial	Final		
A	Acadêmica	2010	janeiro	março	São Paulo	600
B	Privada	2010	janeiro	março	São Paulo	600
C	Privada	2011	janeiro	março	Ceará	501
					Total de laudos coletados	**1701**

Tabela 7: Descrição das amostras coletadas por instituição

3.2 Desenho do estudo

O projeto foi dividido em 2 (duas) fases a saber:

- Fase I: Análise terminológica dos laudos radiológicos;
- Fase II: Avaliação da terminologia por especialistas das instituições envolvidas.

3.2.1 Fase I: Análise terminológica

Na Fase I, os laudos foram comparados para observação de convenções estabelecidas e mapeamento da terminologia comumente empregada. Os termos foram elencados de forma semiautomatizada. Os termos candidatos foram conceituados de acordo com as especificidades da terminologia em língua portuguesa. Nesta fase não houve contato com qualquer colaborador humano, nenhum paciente ou profissional foi contatado, apenas análise dos laudos que são documentos retrospectivos elaborados no primeiro trimestre de 2010 e primeiro trimestre de 2011 foi realizada.

3.2.2 Fase II: Avaliação da terminologia

Na Fase II, foi realizada à avaliação qualitativa da superestrutura textual e terminologia realizada por meio de um questionário com quatro questões objetivas e quatro questões subjetivas totalizando oito questões (vide apêndice B).

Devido à distância entre as instituições, o questionário de avaliação foi construído em suporte eletrônico utilizando a ferramenta de criação de formulários oferecida gratuitamente no Google Drive pela emprega multinacional Google Brasil Internet LTDA. O questionário eletrônico foi enviado por e-mail a radiologistas com pelo menos três anos de experiência em exames de ressonância magnética, portanto, especialistas em ressonância magnética, ligados às instituições que participaram da pesquisa. A explicação da pesquisa e do questionário ocorreu por contato presencial, email ou telefônico, além disso, o termo de consentimento livre e esclarecido (TCLE) foi disponibilizado na primeira parte do questionário eletrônico da pesquisa.

Responderam ao questionário dez radiologistas especialistas distribuídos da seguinte forma: seis da instituição “A”, dois da instituição “B” e dois da instituição “C”. A lista de termos a ser avaliada pelos radiologistas especialistas foi criada utilizando os seguintes critérios:

1. Uma lista de termos que possuíam diferença lexical entre as instituições foi gerada, denominada lista1 contendo 372 (trezentos e setenta e dois) termos;

2. Com base na lista1, o orientador deste trabalho, que é radiologista especialista, excluiu os termos referentes à: anatomia humana (artéria, fígado, etc.), comuns à linguagem natural (correlação, complexo, etc.), descritores de doenças ou patologias de menor importância (cisto, lipoma, etc.) e procedimentos (exemplo: ultrassonografia, tomografia, etc.). Foram excluídos 95 termos e, portanto, a lista2 foi composta por 277 termos.

3. Usando a função “ALEATÓRIOENTRE” do Microsoft Excel um número aleatório foi atribuído a cada um dos 277 termos da lista2 e ordenados do maior para o menor.

4. Após esse processo, foram selecionados os 30 (trinta) primeiros termos no questionário de avaliação.

3.3 Ferramenta computacional *Meta-generalis*

A arquitetura de software empregada para ferramenta computacional *meta-generalis* é de camadas (Chiba e Nardi, 2007), como segue: acesso a dados, regra de negócios, testes unitários (MSDN, 2012) e interface. A figura 17 ilustra a relação entre o banco de dados relacional (RDB) e as camadas da ferramenta. Com a exceção da ferramenta DBDesigner (DBDesigner, 2009) para desenho do modelo de entidade e relacionamento (MER – Model Entity Relation), a suíte de ferramentas Express (gratuitas) da Microsoft foi utilizada, portanto: o RDB utilizado foi o Microsoft SQL Server 2008 R2 Express; a linguagem de programação utilizada foi CSharp (c#); ambiente para desenvolvimento o Visual Studio 2010 Express (MS-VSE, 2012).

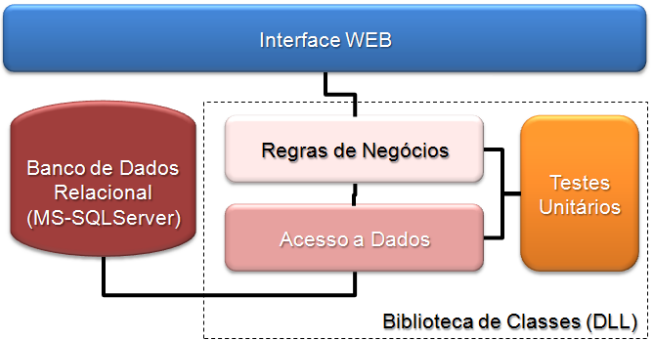


Figura 17: Arquitetura de software da ferramenta computacional *Meta-generalis*.

Detalhes da especificação técnica da ferramenta computacional *Meta-generalis* podem ser obtidas consultando os apêndices D e E.

3.4 Mineração de texto

3.4.1 Aquisição dos laudos radiológicos para formação do corpus textual

Os laudos radiológicos que compuseram as amostras obedeceram aos critérios de seleção, porém foram adquiridos de formas diferentes em cada instituição. Na instituição A, os laudos radiológicos foram mandados em arquivo do programa Microsoft Excel (extensão “xls”) apenas com as partes constituintes descrição e conclusão. No caso da instituição B, o pesquisador principal desta pesquisa foi autorizado a acessar o sistema de informações radiológicas (RIS) onde estavam armazenados os laudos radiológicos que tiveram suas partes constituintes descrição e conclusão transportadas para um arquivo do programa Microsoft Excel (extensão “xls”). A instituição C enviou os laudos na integra em arquivos do Microsoft Word (extensão “doc”) com todas as partes constituintes do laudo incluindo informações do médico solicitante, paciente e médico executante. As autorizações de acesso aos dados podem ser consultadas no anexo A.

3.4.2 Anonimizando dados pessoais

As amostras da instituição C foram enviadas em caráter de confiança e sigilo, mas continham dados pessoais dos indivíduos que participaram da elaboração do laudo, ou seja, o nome completo do paciente dentro do laudo e também no nome do arquivo, o nome completo do médico solicitante dentro do laudo, o nome completo e o número do CRM (Conselho Regional de Medicina) do médico executante (radiologista) dentro do laudo. Frente ao exposto, se fez necessário a dissociação da

relação direta dos indivíduos participantes na elaboração do laudo com seus respectivos dados pessoais.

Para anonimizar os documentos foi preciso substituir as unidades lexicais que representam os dados pessoais por unidades lexicais sem referência a pessoa no mundo real impossibilitando a identificação da pessoa originalmente associada ao laudo. As unidades lexicais substitutas foram extraídas do texto “Finibus Bonorum et Malorum” escrito em latim por possuir palavras de diversos tamanhos e sem significado para o leitor comum (vide quadro seguinte). Este texto é utilizado pela indústria tipográfica desde 1500 d.C. porque cumpre seu objetivo de preencher os espaços textuais sem prender a atenção do leitor (Dena; Douglass; Marino, 2005). Atualmente, ele é encontrado em softwares de distribuição mundial como o Microsoft Word (MS-Support, 2011) e Adobe PageMaker (ADOBE, 2003).

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec ipsum quam, eleifend at luctus quis, viverra eget nisi. Ut vitae diam est, sit amet mattis turpis. Duis eu nibh nec enim elementum fermentum id eu massa. In libero quam, dictum at auctor quis, molestie consectetur elit. Aenean nisi libero, tincidunt cursus facilisis ut, vestibulum eget augue. Sed ipsum nisi, mollis at adipiscing ut, dictum id mauris. Praesent sodales mollis fermentum. Integer purus tellus, gravida quis pellentesque quis, bibendum et mi. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Donec tempus imperdiet nisi eu bibendum. Nulla magna nunc, semper eu adipiscing a, viverra id diam. Phasellus ac sem quis nisl auctor euismod non eget justo. Aliquam arcu risus, tempus eu vestibulum vel, pharetra non velit. Donec diam diam, adipiscing sit amet volutpat a, vehicula a lorem. Donec et nibh odio.

Quadro 1: Um parágrafo do texto Finibus Bonorum et Malorum

Os estudos referentes à instituição C, na ferramenta *Meta-generalis*, foram configurados para passarem pelo processo de anonimização, ou seja, dissociação, descaracterização dos dados pessoais com os laudos da amostra. Ao importar um documento desta instituição o nome do arquivo, o nome do paciente, o nome do

médico solicitante e executante, foi submetido ao método de anonimização que retorna um novo nome (texto) baseado no texto “Finibus Bonorum et Malorum” correspondente ao número de letras de cada unidade lexical do nome, mantendo sua característica textual, sem duplica-lo. Desta forma, um arquivo chamado “Flavio Barbosa.doc” poderia receber o nome de “cursus iaculis.doc”. Embora essa solução funcione para unidades lexicais que representam nomes, para códigos numéricos, como o número do CRM, este método alteraria a função da unidade lexical, por este motivo, ao invés de retornar um nome foi retornado um código sequencial sem qualquer relação com o original. O fluxo completo do processo de anonimização da ferramenta *meta-genaralis* pode ser analisado na figura 18.

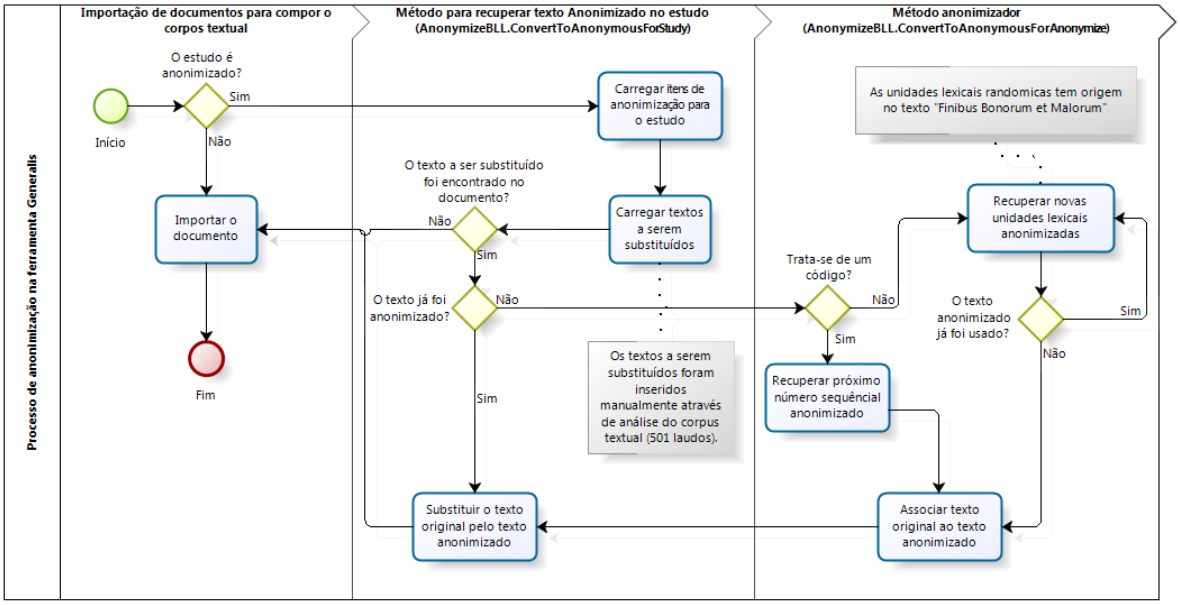


Figura 18: Processo de anonimização na ferramenta *Meta-generalis*

Os nomes e números de CRM a serem anonimizados foram identificados manualmente nos laudos amostrais e inseridos na ferramenta computacional *Meta-generalis*. Após a importação dos documentos para composição do corpus textual a associação entre o texto original e o texto anonimizado pode ser apagada destruindo qualquer relação entre os laudos da amostra (corpus textual) e os seus dados pessoais.

3.4.3 Pré-processamento de importação

O pré-processamento de importação ocorre no momento de importar os arquivos do estudo para a ferramenta computacional *Meta-generalis*. Houve a necessidade de padronizar a tabela de caracteres para ISO 8859-1 Latin 1, pois os arquivos do Microsoft Word eram importados com conjunto de caracteres estranhos.

3.4.4 Pré-processamento do texto para decomposição textual

Antes de efetivamente decompor o texto, foram aplicadas as normalizações de pré-processamento descritas no apêndice F para tornar exato o processamento de decomposição textual.

3.4.5 Processamento do texto para decomposição textual

Durante o processo de decomposição textual as normalizações textuais são aplicadas, por estudo e gramatura com a finalidade de gerar os termos candidatos, são elas: 1) conversão para minúsculo; 2) remoção de acentos; 3) remoção de caracteres especiais e de pontuação; 4) remoção de *stopwords*; 5) conversão de plural para singular; 6) conversão de formas nominais para verbos; 7) conversão de sufixos. Embora a ferramenta computacional *meta-genaralis* permita a geração de termos candidatos com gramatura igual a cinco (nGram=5), neste estudo os resultados são referentes a gramatura “1” (nGram=1) porque os termos compostos foram identificados por decisão ontológica contextual horizontal. Cada tipo de normalização lexical foi acomodado em um estudo independente totalizando sete estudos

terminológicos para cada instituição (vide apêndice F, para maiores detalhes sobre as tabelas de conversão textual por tipo de normalização).

No caso da instituição C, foram consideradas as partes constituintes técnica do exame, descrição do exame, conclusão e, as partes constituintes implícitas a estas. Portanto, as unidades lexicais das partes constituintes a seguir, foram desconsideradas: “identificação do atendimento”, “paciente”, “médico solicitante”, “convênio de saúde”, “médico executante” e “rodapé do laudo radiológico”.

3.4.6 Arquivo invertido

O arquivo invertido foi armazenado no banco de dados relacional utilizando o seguinte conjunto de tabelas: DocumentWord, WordText, TermSource, TermDocument e Term. A descrição completa das tabelas e o MER podem ser consultados no apêndice D.

3.5 Conceitualização

A lista de termos candidatos referente ao corpus textual em estudo foi totalmente conceituada em uma estrutura ontológica, seja de forma automática, utilizando produtos terminológicos eletrônicos, seja manualmente, através da literatura ou consultando especialistas da área.

As relações empregadas entre os conceitos e termos candidatos foram: “É”, para definir, generalizar e ou especializar um conceito (exemplo: espondilose é uma doença degenerativa); “ASSOCIADO”, quando o conceito não é outro conceito, mas existe uma associação entre eles (exemplo: hepático está associado com fígado); “INDICA”, quando determinado termo candidato indica determinado conceito

(exemplo: crânio-vertebral indica uma região anatômica); “PARTE DE”, quando um determinado conceito é parte de outro (exemplo: vértebra é parte da coluna vertebral).

3.5.1 Método de identificação das partes constituintes

Para identificar a superestrutura textual os termos candidatos que possuíam a frequência entre 95% até 100% da quantidade de documentos do estudo foram listados, pois os termos com frequência igual ou próxima a quantidade de documentos do corpus textual são termos candidatos a representarem partes constituintes. Os termos candidatos listados nesta condição foram analisados dentro do contexto que os originou e quando se tratava da identificação de uma categoria textual eles foram relacionados como objetos do conceito “categoria textual” dentro da ontologia na ferramenta computacional *Meta-generalis*.

3.5.2 Método de associação dos termos candidatos as suas partes constituintes de origem

Após a identificação das partes constituintes é possível identificar a exata posição no texto onde a categoria inicia. O fim da categoria textual ocorre quando outra categoria textual se inicia ou o texto termina. Tendo sido determinado o inicio e o fim da categoria textual todos os termos candidatos contidos entre elas foram associados como parte da respectiva categoria textual dentro da ontologia.

3.5.3 Conceituando automaticamente os termos candidatos apoiados por produtos terminológicos eletrônicos

Os termos candidatos foram recuperados do DecS por requisição via protocolo de transferência de hipertexto (HTTP) e interpretação da chave de definição (“<definition>”) dentro do arquivo de linguagem de marcação (XML) retornado. Após a busca por definições dos termos candidatos no DeCS, os conceitos retornados foram avaliados e identificados como compatíveis ou incompatíveis.

A conceitualização de todos os termos candidatos de um determinado estudo, dentro da ferramenta computacional *Meta-generalis*, transforma o próprio estudo em produto terminológico eletrônico (PTE) capaz de retornar o conceito de um determinado termo candidato. Desta maneira, somente o estudo da primeira instituição utilizou apenas o DecS, os demais utilizaram os PTE criado pelos estudos das instituições anteriores.

3.5.4 Conceituando termos candidatos manualmente

Os termos candidatos que não eram conceituados pelos PTE ou não foram adequadamente conceituados por eles ou tiveram de ser conceituados manualmente através da literatura ou consultando especialistas. As fontes literárias consultadas foram: dicionário on-line Michaelis (MICHAELIS, 2009) porque contém o conceito associado ao termo em diversas línguas de especialidade inclusive nas ciências da saúde, o DecS (DecS, 2012), o Dorland dicionário médico ilustrado (Dorland, 1999), o sistema de linguagem médica unificada (UMLS, 2013), artigos científicos, órgãos de pesos e medidas internacionais e atlas de anatomia humana (Netter, 2000).

Para conceituar manualmente os termos candidatos, a ferramenta computacional *Meta-generalis*, exibe a lista de laudos de onde o termo candidato foi extraído o destacando no texto e possibilitando a análise de vários laudos simultaneamente e o significado do termo dentro do seu contexto de origem conforme recomendação do manual de terminologia (Pavel; Nolet, 2002). A figura 19 demonstra como os laudos do termo candidato “disrafismo” foram exibidos dentro do contexto em diversos laudos simultaneamente.

ID.	Arquivo	Descrição
20909	0006-0049.txt	DESCRIÇÃO:Exame realizado em equipamento de 1,5 Tesla, sem a administração endovenosa de contraste. Medula espinhal se estendendo mais inferiormente, junto ao término do saco dural, notando-se lesão ovalada heterogênea, predominantemente com hipersinal T1 e T2, com perda de sinal nas sequências com saturação de gordura, medindo cerca de 2,5 x 1,6 x 1,4 cm, junto à porção terminal da medula lateralizado à esquerda. A lesão apresenta imagens de hpossinal T1 e T2 entremeadas, que podem corresponder a Dilatação do canal medularlombossacro. Inferiormente à lesão descrita, nota-se imagem irregular com hipersinal T1 e T2, com perda de sinal nas sequências com saturação de gordura, em continuidade com a gordura peridural, exteriorizando posterior e lateralmente à esquerda entremeando a musculatura paravertebral (lipomatose?). Coleção no tecido subcutâneo da região sacral à esquerda heterogênea, predominantemente com hipersinal T1 e T2 com imagens de aspecto cístico entremeado, notando-se solução de continuidade da pele nesta região. A coleção mede cerca de 5,7 x 5,6 x 1,5 cm, sem imagem sugestiva de comunicação com o canal vertebral. Alterações em corpos vertebrais lombossacros CONCLUSÃO:1-**Disrafismo** com medula ancorada, lipoma e siringomielia. 2-Coleção subcutânea sugestiva de fístula liquórica Necessário estudo com contraste.
20983	0006-0123.txt	DESCRIÇÃO:Exame realizado em equipamento de 1,5 Tesla. Medula espinhal se estendendo mais inferiormente ao nível da transição lombossacra e término do saco dural, junto a lesão ovalada heterogênea, predominantemente com hipersinal T1 e T2, com perda de sinal nas sequências com saturação de gordura, medindo cerca de 2,5 x 1,6 x 2,2 cm. Dilatação do canal medular lombossacro. Escoliose com dupla curvatura, apresentando alteração da morfologia das vértebras torácicas e lombossacras, com hemivértebras tóracicas e lombares e falha de fechamento dos elementos posteriores lombares e sacrais. Assimetria do corpo vertebral de L2, mais largo à esquerda. CONCLUSÃO:1-**Disrafismo** com medula ancorada, lipoma e siringomielia. 2-Escoliose de dupla curvatura com presença de hemivértebras.
21232	0006-0372.txt	DESCRIÇÃO:Intensidade de sinal normal da medula óssea dos corpos vertebrais lombares e torácicos baixos, da porção terminal da medula espinhal e cone medular, das raízes, líquor, núcleos pulposos e ânulos fibrosos dos discos vertebrais estudados. Ausência de saliências discais. Não há sinais de compressão medular ou radicular. Dimensões normais do canal vertebral e recessos laterais. Discretas alterações degenerativas em articulações facetárias. Gordura epidural e musculatura paravertebral sem alterações.CONCLUSÃO:Discretas alterações degenerativas em articulações facetárias. Exame sem evidências de **disrafismo** oculto.
21276	0006-0416.txt	DESCRIÇÃO:Intensidade de sinal da medula óssea normal. Ainda não se observa a fusão óssea completa dos arcos neurais posteriores das vértebras L5 e S1. A fusão completa dos núcleos de ossificação em geral se dá entre 1 e 7 anos de vida. O cone medular e as raízes da cauda equina se apresentam com volume, morfologia e sinal dentro da normalidade. O término do cone medular está situado na altura da vértebra L1. Não foram identificadas lesões expansivas ou sinais de **disrafismo**s ocultos. Os núcleos pulposos e ânulos fibrosos dos discos vertebrais estudados se apresentam de aspecto normal. Ausência de hérnias discais. Não há sinais de compressão medular ou radicular. Dimensões normais do canal vertebral e recessos laterais. Intensidade normal das facetas articulares. Gordura epidural e musculatura paravertebral sem alterações. Observação: A avaliação de resíduo vesical e da capacidade vesical não é realizada por meio da resonância magnética da coluna lombar. A critério clínico sugiro avaliação por meio da ultrassonografia.CONCLUSÃO:Exame de coluna lomobossacra sem anormalidades significativas.

Figura 19: Visualizando os laudos de onde o termo candidato "disrafismo" foi extraído para ser conceituado

3.6 Aspectos éticos

O estudo foi submetido à apreciação do Comitê de Ética em Pesquisa do HCFMRP-USP, através da Plataforma Brasil do Ministério da Saúde e, foi aprovado por este comitê em 16/07/2012 sob o número CAAE 00593812.2.1001.5440. Portanto, o TCLE foi dispensado na fase I por se tratar de laudos retroativos anônimos ou anonimizados. Na fase II, os radiologistas especialistas que participaram como avaliadores aderiram ao TCLE (vide apêndice A).

4 RESULTADOS

4.1 Quantitativo

4.1.1 Normalização

A quantidade de unidades lexicais brutas extraídas das instituições A, B e C foi respectivamente de 2633, 1680 e 1816 totalizando 6129. Após o sétimo e, último, tipo de normalização a quantidade de unidades lexicais normalizadas nas instituições A, B e C foi respectivamente de 1949, 1193 e 1426 totalizando 4568.

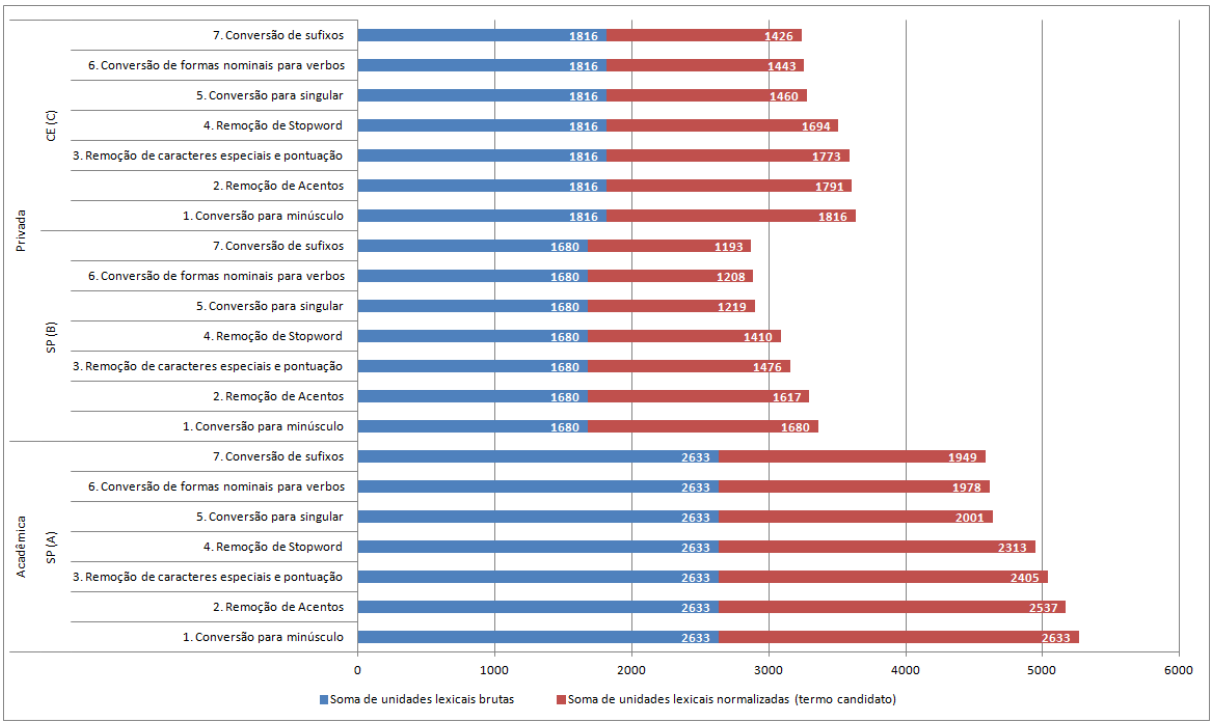


Gráfico 3: Resultado absoluto das normalizações por instituição

Para comparar o efeito de cada tipo de normalização entre as instituições utilizou-se a densidade léxica (DL), aglutinação léxica (AL) e a diferença da aglutinação léxica (dAL) com base no tipo de normalização anterior, conforme demonstra a tabela 8:

Tipos de Normalização	Instituição A			Instituição B			Instituição C		
	DL	AL	dAL	DL	AL	dAL	DL	AL	dAL
1 Conversão para minúsculo	100,0%	0,0%	0,0%	100,0%	0,0%	0,0%	100,0%	0,0%	0,0%
2 Remoção de Acentos	96,4%	3,6%	3,6%	96,3%	3,8%	3,7%	98,6%	1,4%	1,4%
3 Remoção de caracteres especiais e pontuação	91,3%	8,7%	5,1%	87,9%	12,1%	8,4%	97,6%	2,4%	1,0%
4 Remoção de Stopword	87,8%	12,2%	3,5%	83,9%	16,1%	4,0%	93,3%	6,7%	4,4%
5 Conversão para singular	76,0%	24,0%	11,8%	72,6%	27,4%	11,3%	80,4%	19,6%	12,9%
6 Conversão de formas nominais para verbos	75,1%	24,9%	0,9%	71,9%	28,1%	0,7%	79,5%	20,5%	0,9%
7 Conversão de sufixos	**74,0%**	**26,0%**	1,1%	**71,0%**	**29,0%**	0,9%	**78,5%**	**21,5%**	0,9%

Tabela 8: Densidade léxica (DL), aglutinação léxica (AL) e a diferença da aglutinação léxica (dAL) por instituição e tipo de normalização

Os gráficos 4 e 5 foram construídos com base na tabela 8, porém para facilitar o entendimento, o primeiro compara a densidade léxica (DL) com a aglutinação léxica (AL) e o segundo a densidade léxica (DL) com a diferença de aglutinação léxica (dAL), ambos por instituição e tipo de normalização léxica.

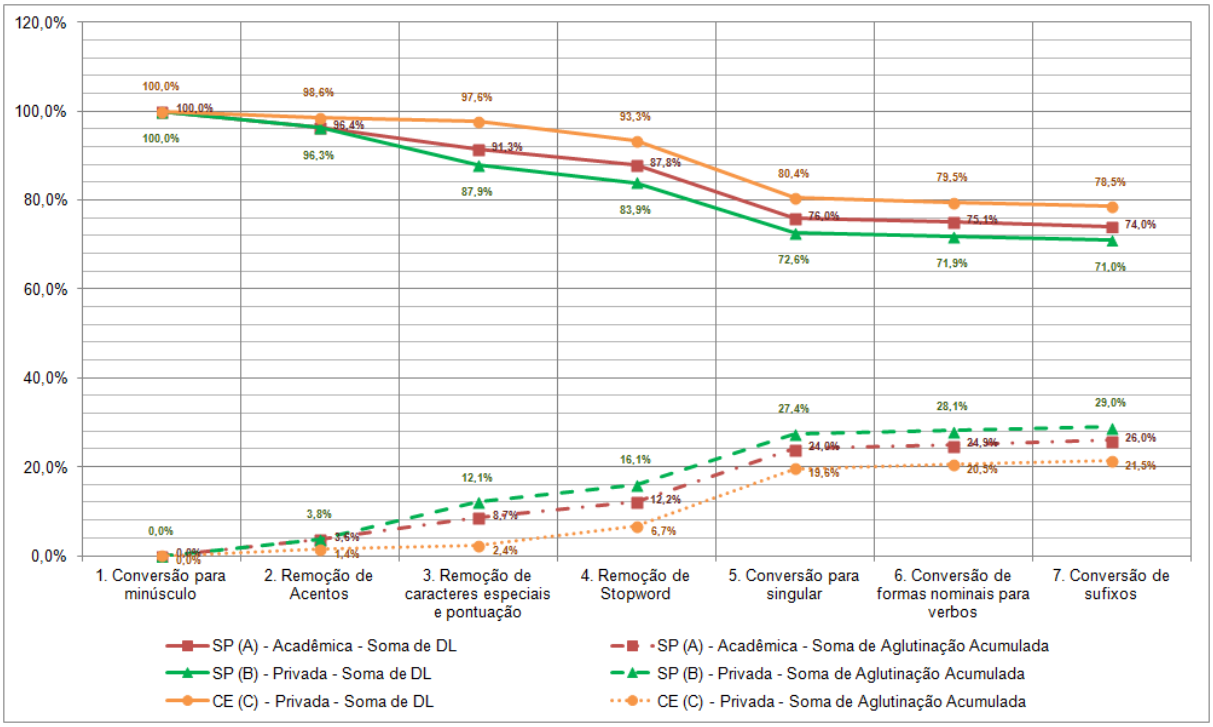


Gráfico 4: Densidade léxica (DL) e aglutinação léxica (AL) por instituição e tipo de normalização

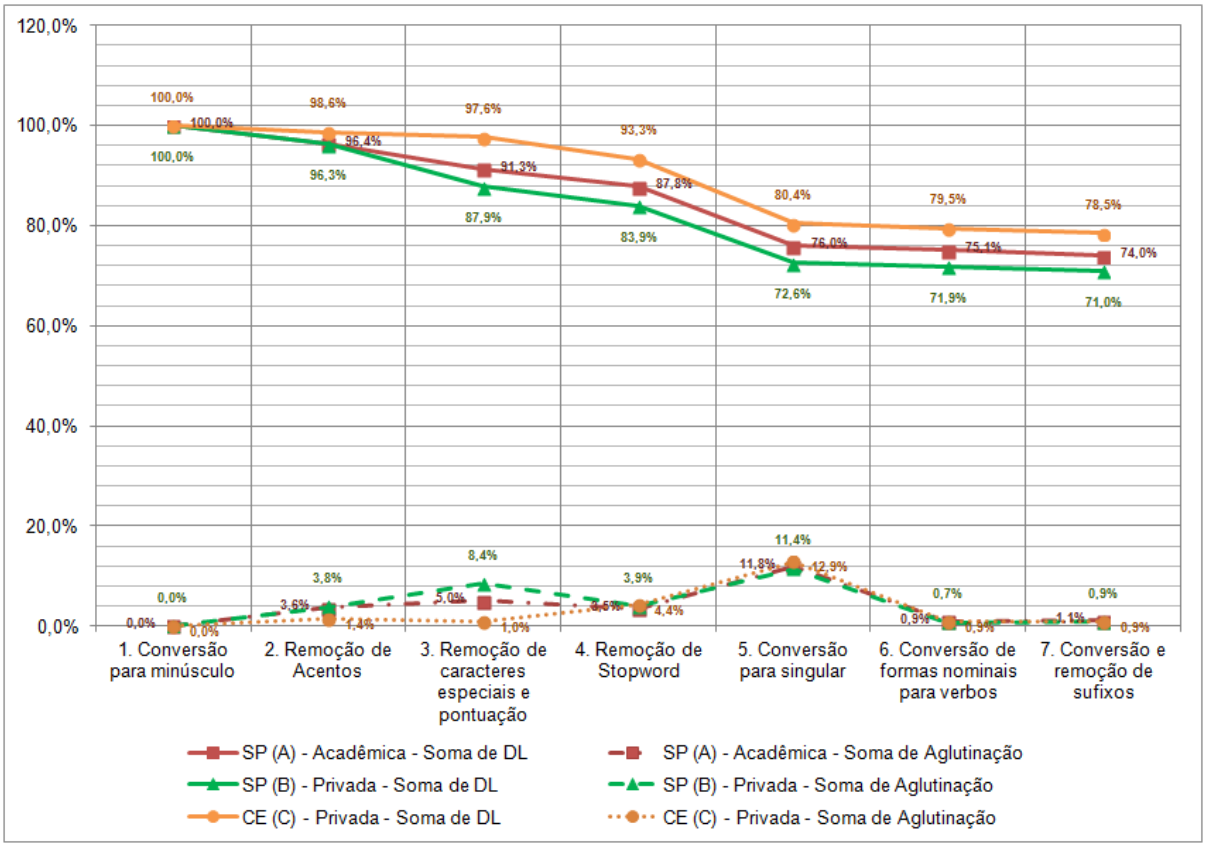


Gráfico 5: Densidade léxica (DL) e diferença de aglutinação léxica (dAL) por instituição e tipo de normalização

Após aplicação do sétimo e último tipo de normalização a densidade léxica (DL) das instituições A, B e C foi respectivamente de 74,0%, 71,0% e 78,5% e a aglutinação léxica (AL) de 26,0%, 29,0% e 21,5% (vide gráfico 5).

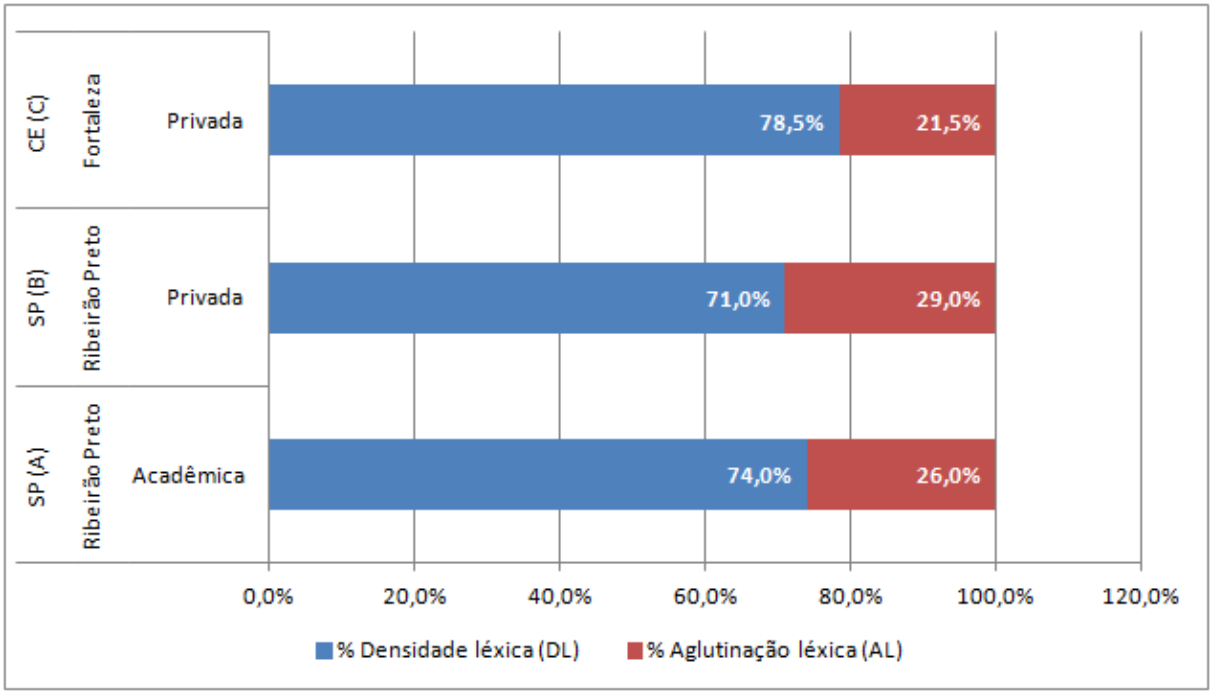


Gráfico 6: Densidade léxica (DL) e aglutinação léxica (AL) das instituições após a última normalização

4.1.2 Superestruturas textuais identificadas

Foram identificadas três superestruturas textuais do laudo radiológico, uma para cada instituição, contendo cinco partes constituintes, com variação terminológica em suas denominações, a saber: “Técnica do Exame” (explícita), “Descrição”, “Técnica do Exame no primeiro parágrafo da descrição” (implícita), “Conclusão”, e “Observação”.

A superestrutura textual do laudo radiológico da instituição A é composto pelas partes constituintes “Descrição”, “Conclusão” e “Técnica do Exame” no primeiro parágrafo da “Descrição” (implícita), porém, esta última, foi utilizada em 19% dos laudos radiológicos da amostra.

Na instituição B, foram identificadas as partes constituintes “Descrição”, “Impressão diagnóstica” e “Técnica do Exame” no primeiro parágrafo da “Descrição” (implícita). A categoria textual “Impressão Diagnóstica” foi identificada pela abreviatura

"I.D." e foi utilizada para informar de maneira conclusiva as observações da imagem, assim como, a categoria textual "Conclusão" identificada na instituição A.

No caso da instituição C, foram identificadas as partes constituintes "Resultado", "Conclusão", "Técnica do Exame" e "Observação", todas as partes constituintes independentes no texto (explícitas). A categoria textual "Resultado" foi utilizada para descrever os achados da imagem, assim como, a categoria textual "Descrição" identificada na instituição A e B. Observou-se que, para a categoria textual "Conclusão", o termo "Impressão Diagnóstica" foi utilizado em 5% dos laudos da amostra desta instituição, ou seja, ocorreu uma variação terminológica, um caso de sinonímia. A categoria textual "Observação" foi identificada pela abreviatura "OBS:" e ocorreu em 8,5% dos laudos da amostra.

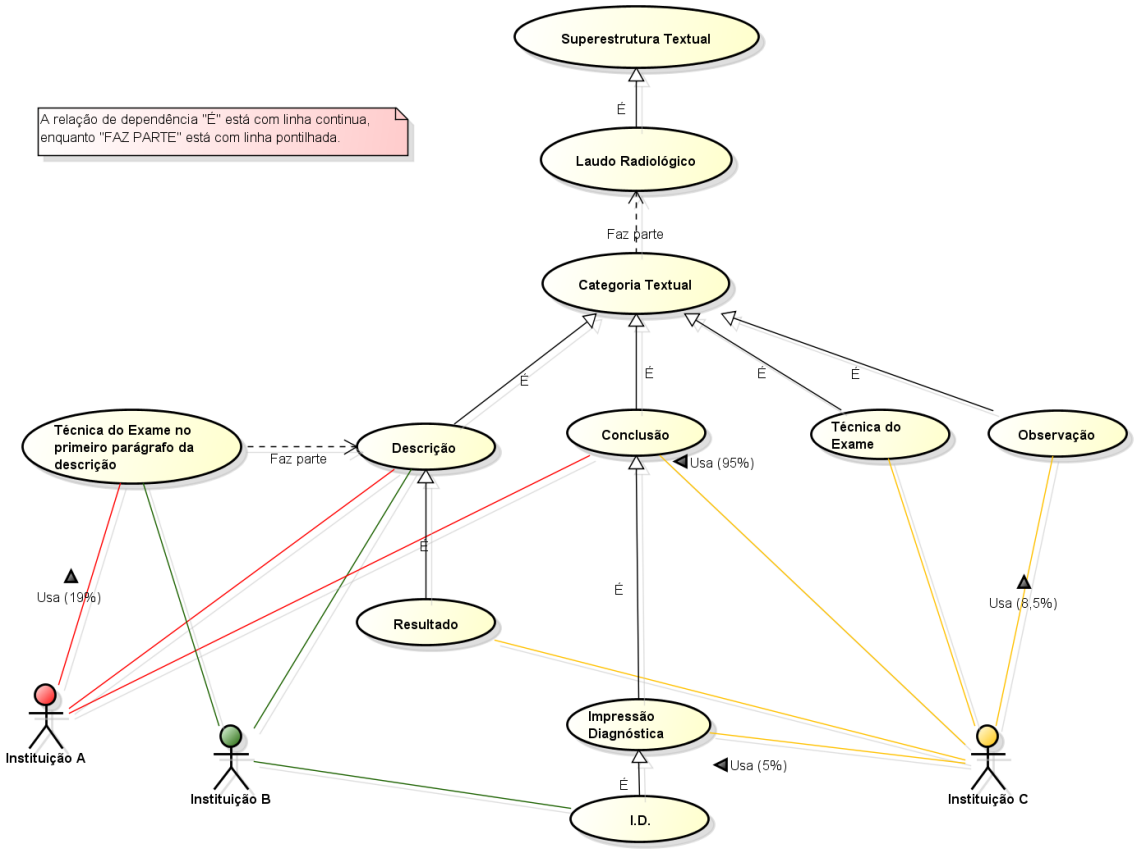


Figura 20: Esquema de utilização das superestruturas textuais identificadas nas instituições

4.1.3 Conceitualização

Durante o processo de conceitualização, 4568 termos candidatos foram analisados e, destes, 4183 termos candidatos foram conceituados originando 727 conceitos únicos em português. As instituições A, B e C compartilham dos conceitos utilizando respectivamente 606, 387 e 431 conceitos (vide tabela 9).

Instituição	Termos Candidatos	Conceitos	Conceitos exclusivos da instituição	% Termos candidatos conceituados	% Termos candidatos conceituados exclusivamente
A	1949	606	185	31%	9%
B	1193	387	38	32%	3%
C	1426	431	67	30%	5%

Tabela 9: Síntese dos conceitos gerados por instituição

Os termos candidatos estão relacionados com os conceitos por 13963 relações onde 6018 relações são usadas na instituição A, 2875 são usadas na instituição B e 5070 são usadas na instituição C. Devido ao compartilhamento dos conceitos entre as instituições a sua distribuição não é uniforme.

Instituição	A		B		C		Considerando todas as instituições	
Relação	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito
Associado	7	104	5	82	22	86	34	105
É	1548	250	939	220	998	234	3485	262
Indicado	596	287	367	209	293	193	1256	340
Parte de	3867	80	1564	55	3757	64	9188	82
	6018	721	2875	566	5070	577	13963	789

Tabela 10: Síntese das relações entre termos e conceitos

A primeira instituição que passou pelo processo de conceitualização dos termos candidatos foi a instituição A por possuir a maior quantidade de unidades lexicais

normalizadas. Quando os termos candidatos desta instituição foram submetidos de forma automática ao DecS, foram conceituados 22,1% de forma automática e 77,9% manualmente.

A instituição B teve seus termos candidatos submetida de forma automática ao produto terminológico eletrônico (PTE) criado pelo estudo da instituição A e, foram conceituados automaticamente 77,8% dos termos candidatos e 22,2% manualmente.

Por fim, a instituição C, teve seus termos candidatos submetidos de forma automática ao PTE criado pelo estudo da instituição A e B e, foram conceituados automaticamente 83,1% dos termos candidatos e 16,9% manualmente.

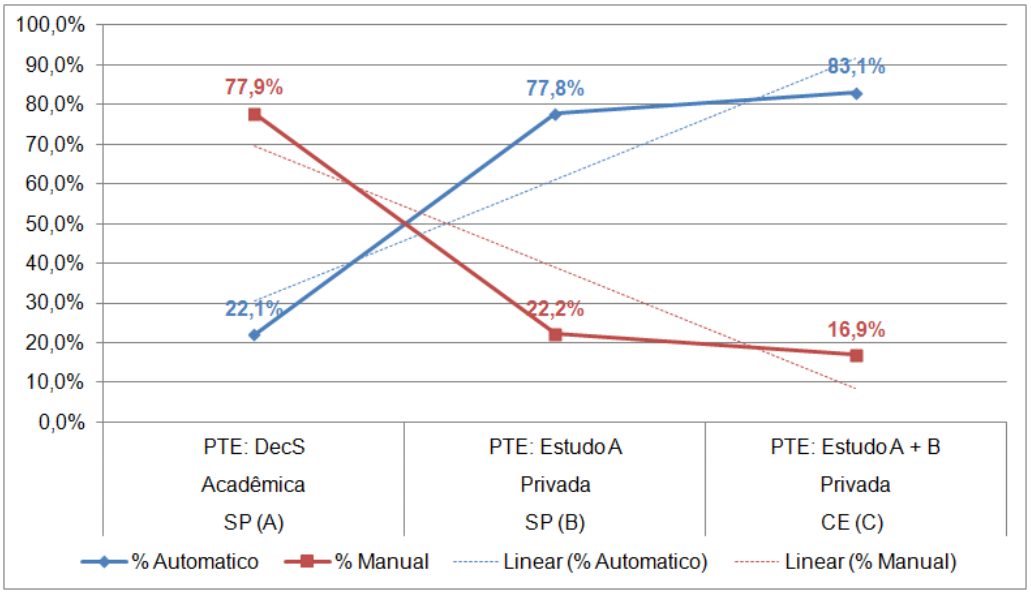


Gráfico 7: Proporção de termos candidatos conceituada manualmente e automaticamente por produto terminológico eletrônico (PTE) e instituição

Duas linhas de tendência foram traçadas sobre o gráfico da proporção de termos candidatos conceituada manualmente e automaticamente por PTE e instituição (vide gráfico 7) para evidenciar a tendência de menor necessidade de conceitualização manual quanto maior a quantidade de PTE utilizados, principalmente os extraídos da própria linguagem de especialidade em estudo.

4.2 Qualitativo

Por se estar propondo uma metodologia, o questionário utilizado é um instrumento exploratório com a finalidade de avaliar a superestrutura e terminologia extraída utilizando a metodologia apresentada. O questionário completo pode ser consultado no apêndice B.

O questionário de avaliação enviado aos dez (n=10) radiologistas especialistas das instituições participantes desta pesquisa não os identificou e, portanto, todas as respostas são anônimas. Todavia, cada avaliador recebeu um código de identificação sequencial de modo que o primeiro é o A1 e o último A10.

4.2.1 Sobre a superestrutura

Para avaliar as superestruturas identificadas nos laudos do estudo foram apresentadas três superestruturas textuais nomeadas da seguinte forma: “Modelo A”, “Modelo B” e “Modelo C”.

A superestrutura do “Modelo A” foi composta por duas partes constituintes denominadas respectivamente “Descrição” e “Conclusão”, onde, dentro da “Descrição” poderia existir uma categoria textual chamada “Observações” a ser preenchida conforme a necessidade. No “Modelo B”, a superestrutura foi constituída por duas partes constituintes explícitas, denominadas respectivamente por “Descrição do laudo” e “Impressão diagnóstica” e, uma categoria textual implícita denominada “Técnica do exame” que é informada no primeiro parágrafo da categoria textual “Descrição do laudo”. Por último, o “Modelo C”, foi constituído de três partes

constituintes explicitas denominada respectivamente “Técnica do Exame”, “Descrição” e “Resultado”.

Os modelos apresentados estão baseados nas superestruturas observadas na amostra, porém eles não são exatamente equivalentes nas três instituições, modificando por vezes, a posição das partes constituintes e o termo que descreve cada categoria textual aderente a superestrutura.

4.2.1.1 Questão 1: Qual a estrutura de laudo mais adequada?

Modelo	Quantidade de respostas
A	0
B	6
C	4

Tabela 11: Tabulação das respostas dos avaliadores à questão 1 (n=10).

Nenhum avaliador escolheu o “Modelo A”, conforme observado nas justificativas apresentadas na questão dois, porque não havia no “Modelo A” informação sobre a técnica do exame e o termo “Impressão Diagnóstica” é mais apropriado que os termos “Conclusão” ou “Resultado”.

A maioria dos avaliadores prefere o “Modelo B” porque nele se encontram as partes constituintes essenciais para confecção do laudo radiológico (Prestes-Júnior; Tourinho; Rangel, 2012) e o termo “Impressão Diagnóstica” ao invés de “Conclusão” ou “Resultado”.

A porção de avaliadores que escolheram o “Modelo C”, o fez por considerar mais importante explicitar a categoria textual “Técnica do exame”, do que o termo utilizado para representar a categoria textual conclusiva do laudo radiológico, seja ele, “Impressão diagnóstica”, “Resultado” ou “Conclusão”.

4.2.1.2 Questão 2: Comente sobre a sua escolha do modelo.

Avaliador	Modelo Escolhido	Resposta
A1	B	Creio que não há uma forma única e correta para relatórios. Em parte a escolha está relacionada a preferências pessoais. Mas há quesitos que são essenciais, como uma parte descritiva e outra na qual o radiologista dá sua opinião baseada em sua experiência e em seu conhecimento de especialista. Eu prefiro que esta última parte se denomine impressão diagnóstica embora não veja como errôneo chama-la de parte conclusiva. Os modelos apresentados me parecem diferir inicialmente mas, na prática podem ser muito semelhantes.
A2	C	Acho que facilita para o médico solicitante visualizar as informações que seriam mais importantes para ele.
A3	B	Sequência mais lógica.
A4	C	Composta da sequencia do tipo de exame; descrição dos achados e a impressão diagnóstica de acordo com a suspeita e achados.
A5	C	Acho que fica mais claro quando se separa a técnica do exame do corpo do laudo. Alem disso acho que o laudo deve ter uma parte descritiva e outra conclusiva.
A6	C	É o modelo que uso.
A7	B	Considero a descrição da técnica do exame importante. Considero que a síntese de achados ou a interpretação do medico Radiologista deve ser denominada impressão diagnostica.
A8	B	Acho a sua estruturação mais adequada, pois apresenta uma impressão diagnóstica baseada na interpretação das imagens e não um "resultado". O modelo A também se aproxima do que acho ideal, mas não se refere à técnica do exame.
A9	B	Ainda na alternativa B, a escolhida, faria observações quando necessário, inclusive sobre a qualidade técnica das imagens.
A10	B	Acredito que o laudo completo deva conter os aspectos técnicos da realização do exame, de preferência em primeiro plano, para já adiantar possíveis problemas ou restrições do exame para a adequada avaliação das imagens. Descrição pormenorizada dos achados e conclusão me parece a sequencia lógica. Creio que o termo impressão diagnostica seja o termo que melhor exprima a avaliação global do exame disposta de forma mais concisa e objetiva dos achados, associada à minha interpretação

	pessoal no momento de responder aos questionamentos do medico solicitante.

Tabela 12: Tabulação das respostas dos avaliadores à questão 2 (n=10).

Mais da metade dos avaliadores preferem o termo “Impressão Diagnóstica” ao invés do termo “Resultado” contando os que escolheram o “Modelo B” e o avaliador A4 que escolheu o “Modelo C”, porém retifica o termo ao escrever “Impressão diagnóstica” ao invés de “Resultado” em sua justificativa da opção por este modelo.

4.2.1.3 Questão 3: Você sugeriria outra estrutura para o modelo de laudo da coluna lombo sacra? Descreva abaixo. (Se deixar em branco entenderemos que o modelo escolhido é perfeito para elaboração do laudo radiológico da coluna lombossacra).

Avaliador	Resposta
A4	Acrescentaria os dados clínicos no início do relatório.
A7	Sugiro acrescentar o espaço para observações, quando necessárias do modelo A.
A8	Adotaria o modelo B. Mas como o modelo A contempla espaço para observações eu também incluiria essa possibilidade no modelo B e utilizaria esse espaço para incluir alterações de outras estruturas que possa ser parcialmente visualizadas no exame e para comparação entre laudos.
A10	Não vejo necessidade de laudo especifico e diferenciado para coluna lombossacra.

Tabela 13: Tabulação das respostas dos avaliadores à questão 3 (n=10).

Menos da metade dos avaliadores responderam a esta questão. Entre os que responderam, nenhum sugeriu uma nova informação que devesse ser preenchida pelo radiologista ou uma nova superestrutura textual.

Os dados clínicos, sugerido pelo avaliador A4, fazem parte da solicitação do exame e é preenchido pelo medico solicitante. Geralmente o laudo radiológico é escrito acoplado ao sistema de informações radiológicas (RIS, Radiology Information System) que recupera as informações clinicas inseridas pelo medico solicitante e as disponibiliza para o medico executante, no caso, o radiologista. Os avaliadores A7 e A8 sugerem que a categoria textual para informação de observação deva ser incluída no modelo que escolheram e ser preenchida quando necessário. Do ponto de vista do avaliador A10, a superestrutura apresentada é suficientemente abrangente para contemplar o laudo de ressonância magnética da coluna lombossacra.

4.2.2 Sobre a terminologia

Por se tratar de uma pesquisa exploratória, alguns ensaios com um instrumento eletrônico que agrupando as informações em arvore e, chegando ao nível do laudo radiológico anonimizado foram avaliadas, por fim, conforme descrito na metodologia, os termos foram aleatórios e escolhido segundo alguns critérios. É possível consultar os termos disponibilizados pelos avaliadores no apêndice B.

4.2.2.1 Questão 4: Os termos apresentados são adequados para a descrição de achados?

Resposta	Quantidade de respostas
Concordo plenamente	4
Concordo parcialmente	6
Discordo plenamente	0
Discordo parcialmente	0
Não posso avaliar	0

Tabela 14: Tabulação das respostas dos avaliadores à questão 4 (n=10).

Nenhum avaliador discordou que os termos são adequados e todos puderam avaliar os termos. Todavia, a maioria dos avaliadores discorda da definição de alguns termos como será possível constatar nas respostas da questão número seis.

4.2.2.2 Questão 5: Qual a quantidade de termos apresentados que é empregada na sua instituição para elaboração do laudo?

Opções	Quantidade de respostas
Nenhum dos termos apresentados	0
1 e 8 termos	0
9 e 15 termos	0
16 e 23 termos	3
24 e 30 termos	7

Tabela 15: Tabulação das respostas dos avaliadores à questão 5 (n=10).

Em consequência ao sorteamento dos termos a serem apresentados aos avaliadores, a proporção de termos entre as instituições não ficou uniforme, sendo

distribuída pelas proporções: 100% dos termos existiam na instituição A, 60% dos termos (18 termos) existiam na instituição B e C (vide apêndice C).

O resultado esperado, considerando a distribuição dos termos entre as instituições, seria 40% dos avaliadores tivessem escolhido entre 16 e 23 termos.

4.2.2.3 Questão 6: Dos termos e conceitos apresentados, existe algum que não corresponde ao significado empregado na sua instituição (clínica)? Qual ou quais?

Avaliador	Resposta
A1	Espondilose.
A3	Campo visual e espondilose.
A5	Hidrossiringomielia: Ha uma diferença entre hidromielia e siringomielia em termos conceituais, porem como não se pode diferencia-las pala RM se conclui como uma junção dos termos, porém o conceito aqui apresentado é o de hidromielia, ou seja, especificamente a dilatação do canal central da medula, enquanto a seringomielia refere-se a qualquer cavidade cística intramedular fora do cana central da medula.
A9	Hidrossiringomielia.

Tabela 16: Tabulação das respostas dos avaliadores à questão 6 (n=10).

Dos trinta termos apresentados, quatro foram considerados não correspondentes ao significado empregado na instituição do avaliador.

4.2.2.4 Questão 7: Facilitaria a emissão do laudo radiológico a existência de um vocabulário comum entre diferentes instituições.

Opções	Quantidade de respostas
Concordo plenamente	6
Concordo parcialmente	4
Discordo plenamente	0
Discordo parcialmente	0
Não posso avaliar	0

Tabela 17: Tabulação das respostas dos avaliadores à questão 7 (n=10).

Embora, alguns avaliadores não concordem plenamente, a maioria deles está de pleno acordo que um produto terminológico comum entre diversas instituições facilitaria a emissão do laudo radiológico. Nenhum dos avaliadores discordou que existe facilidade em um produto terminológico comum a várias instituições.

4.2.2.5 Questão 8: Por favor, se preciso, adicione seus comentários.

Avaliador	Resposta
A4	A padronização do laudo apresenta benefícios para o paciente em primeiro lugar, por padronizar também a conduta e assim contribuir para um melhor resultado do tratamento.
A5	Excelente iniciativa. Importantíssimo tentar homogeneizar terminologia em laudos de coluna que são particularmente heterogêneas. Porém senti falta da terminologia usada para alterações em discos intervertebrais como hérnias, abaulamentos protrusões etc., que apesar de existir um consenso publicado ainda e objeto de muita discussão.
A5	Acredito que uma uniformização de laudo e termos seja benéfico para o trafego de informações entre o paciente e diferentes médicos ou então mesmo numa conversa de bar entre o paciente e amigos. No entanto, creio que seja de pouca relevância dentro do contexto medico vide que embora exista pluralidade de termos e expressões, o entendimento dos diversos laudos por diferentes médicos se faz sem muitas dificuldades. Talvez seja possível o desenvolvimento de ferramentas que se utilizam de laudos uniformes no intuito de facilitar a confecção de laudos, reduzindo o tempo necessário para sua elaboração. E isso pode realmente funcionar em alguns casos porem eu acredito que cada paciente é um paciente e cada exame é um exame, independente das semelhanças que existam. Sabe-se que os mesmos achados de hérnias discais, por exemplo, podem ter significados clínicos completamente diferentes. O radiologista deve sempre saber relevar os achados dos exames, com base na clinica do paciente fornecida pelo medico solicitante e nos questionamentos do solicitante que motivaram a realização do exame. Isso se dá de forma mais pratica com um texto livre e particularizado para cada exame. Por fim, gostaria de salientar que no meu entendimento, os laudos radiológicos são diálogos entre médicos. Nesse diálogo, o radiologista obrigatoriamente deve ter a liberdade para mudar as estruturas de laudo quando isso se faz necessário, a fim de fazer as suas analises e interpretação do exame prevalecerem de forma mais clara e nítida. A pratica médica, incluindo a radiologia, é também uma espécie de arte, aonde é difícil ser impessoal e robotizado. "Acima de todas as liberdades, dê-me a de saber, de me expressar, de debater com autonomia, de acordo com minha consciência." John Milton.

Tabela 18: Tabulação das respostas dos avaliadores à questão 8 (n=10).

4.3 Metodologia meta-generalis

A metodologia *meta-generalis* defendida por este trabalho é constituída por seis passos: 1) coleta dos laudos radiológicos de determinado tipo de exame; 2) decomposição textual; 3) normalização das unidades lexicais; 4) identificação das superestruturas textuais; 5) conceituação dos termos candidatos; 6) avaliação das superestruturas identificadas e da terminologia extraída pelos especialistas na linguagem de especialidade estudada.

5 DISCUSSÃO

O processo de construção da ferramenta computacional *meta-genaralis* foi motivado, principalmente, por não existir uma ferramenta capaz de minerar textos em português e, de forma integrada, permitir a conceituação contextualizada dos termos candidatos em estrutura ontológica armazenada em RDB. A maior parte dos programas fazem uma coisa ou outra, ou mineram textos ou criam ontologias e, raramente utilizam RDB, que permite inferências a qualquer momento, sem necessidade de reprocessamento para carga dos dados em memória. Além disso, nenhum programa que considere ou identifique as superestruturas textuais presentes no corpus textual foi encontrado.

Modelar uma estrutura de dados capaz de acomodar a ficha terminológica (Pavel; Nolet, 2002) por idioma e formalizá-la ontologicamente, criar algoritmos na linguagem de programação C# para contemplar os arquivos invertidos, conversão textual, anonimização, gravação em RDB, foram desafios técnicos interessantes desta pesquisa.

Diversos autores entendem que a explicitação do conhecimento passa invariavelmente pela conceitualização que é possível através da formalização de ontologias (Schiessl & Bräscher, 2011). Este estudo corrobora com estes autores ao encontrar avaliadores que identificaram no termo “espondilose” um erro conceitual, quando na verdade, eles possuíam um conceito incompleto sobre o objeto que este termo representa no mundo real. Portanto, é um fato que o conhecimento não está nas unidades lexicais (palavras), mas sim, dentro de cada membro de uma comunidade especializada. Nomear agrupamento de unidades lexicais frequentes e mostrar os documentos onde elas foram encontradas indica a temática do corpus textual e propicia a recuperação de informação, mas isso não transforma as unidades lexicais em conhecimento explícito. A variação terminológica existe porque o

profissional pode formular um modelo mental, um conceito, equivocado ou incompleto sobre qual o objeto que está sendo referenciado no mundo real pelo termo que o representa. Isto pode ocorrer por fatores linguísticos e culturais (por exemplo, quando um conceito é referenciado por dois termos diferentes, um em inglês e outro em português) ou por conta de sua formação e experiência profissional (por exemplo, quando o profissional identifica um termo e mentaliza um conceito incompleto sobre o objeto do mundo real que ele representa). Portanto, o conhecimento estará formalizado quando o sistema de informação proporcionar entendimento unívoco sobre um determinado objeto do mundo real em uma comunidade especializada (contexto).

Um problema frequentemente enfrentado na análise de corpus textuais e na investigação cientifica de modo geral é a necessidade de descaracterizar, anonimizar, a identificação das pessoas envolvidas na pesquisa (Martins & et. al., 2013). Tanto a substituição da unidade lexical, como a sua remoção, influenciam diretamente as estatísticas do documento na mineração de texto e pode inviabilizar estudos que visam acompanhar a evolução de uma determinada variável ao longo dos textos para um determinado individuo, por exemplo, a quantidade de nódulos que o paciente “X” apresentou no corpus textual “Y”. A remoção dos dados pessoais dos documentos dificulta a análise manual realizada por humanos porque cria espaços vazios no texto (buracos) e comprometem sua função comunicativa. Para equacionar essa problemática, a técnica de anonimização dos dados de identificação do indivíduo descrita nesse trabalho adotou a substituição do nome do paciente ou do profissional por unidades lexicais baseadas no texto “Finibus Bonorum et Malorum” ou código identificador (CRM, por exemplo), por número sequencial, que se mantém durante a análise do corpus textual com a mesma quantidade de caracteres do número

identificador original, preservando assim, semelhança das unidades lexicais substituídas com suas respectivas substitutas.

Os sete tipos de normalização conseguiram aglutinar o corpus textual composto pelos laudos radiológicos das três instituições estudadas em média 25,5% conservando as unidades lexicais inteligíveis. O primeiro tipo de normalização, conversão para minúsculo, obviamente não aglutinou o texto, porém permitiu ao computador realizar comparações exatas aumentando a velocidade de processamento, pois não havia preocupação com a diferença entre maiúsculos e minúsculos. O tipo de normalização com maior aglutinação entre todas as instituições foi a transformação do plural para o singular, em média 12%. Houve pico de 8% na remoção dos caracteres especiais e pontuação na instituição B porque os textos usavam caracteres como hífen de marcadores. No demais, a normalização teve igual efeito entre as instituições (vide tabela 8 e gráfico 4 e 5). No total, foram normalizadas 6129 unidades lexicais obtendo 4568 termos candidatos dos quais 4183 foram conceituados e originaram 727 conceitos.

A superestrutura textual de laudos radiológicos nas três instituições foi semelhante, porém observou-se variação terminológica (sinonímia) entre os nomes das partes constituintes de uma mesma instituição e entre instituições diferentes. Na instituição A e B a categoria textual destinada à descrição dos achados clínicos denominava-se "Descrição", enquanto na instituição C era "Resultado".

A instituição C apresentou maior organização das partes constituintes as deixando explicitas, destacado do texto, independente das outras partes constituintes como a "Técnica do Exame" que na instituição A e B estava implícita, embutida, no primeiro parágrafo da categoria textual "Descrição". Todavia, na instituição C, em 5% dos laudos continham o termo "Impressão Diagnóstica" ao invés de "Conclusão" na

categoria textual destinada a este fim (vide figura 20) e, embora a instituição B, assim como a instituição C, utilizava-se da categoria textual nomeada por “Impressão Diagnóstica” ela estava abreviada como “I.D.” caracterizando outro caso de sinonímia.

A instituição A usou a técnica do exame no primeiro parágrafo em 19% dos seus laudos, provavelmente nos casos em que eram necessárias informações adicionais, visto que, o RIS desta instituição possui pedido de exame eletrônico com superestrutura, contendo partes constituintes com descritores pré-definidos da região, exame e a técnica solicitada.

O processo de conceitualização é o mais trabalhoso para construção do PTE, porque embora exista a conceituação automática os termos candidatos precisavam ser analisados dentro do seu contexto a fim de verificar a compatibilidade do conceito associado e a unidade lexical normalizada. Foram criados conceitos que representavam classes de objetos, por exemplo “vértebra lombar”, e os objetos destas classes relacionados a elas, exemplo “L1” é uma “vértebra lombar”. Além de relacionar termos candidatos aos respectivos conceitos a estrutura ontológica permitiu relacionar conceitos a outros conceitos formalizando relações que não estavam explicitamente na terminologia extraída do corpus textual, mas sim, na relação e classificação do objeto no mundo real. Por exemplo, ao empregar o termo “espondilose”, o radiologista, se refere a um problema especifico que acomete a coluna vertebral, todavia, “espondilose” pode ser classificada de forma geral como uma “doença”, desta forma, empregando a relação “é”, relaciona-se o conceito “espondilose” com o conceito que representa sua classe geral “doença”. A lista de termos candidatos referente ao corpus textual das três instituições foi totalmente conceitualizada, estabelecendo 13963 relações entre termos candidatos e conceitos, e 789 relações entre conceitos.

Utilizar PTE específicos da linguagem de especialidade permite uma maior conceituação, porque a diferença entre a recuperação automática de conceitos do DecS e o PTE das instituição A e B foi 61% maior. Não houve grandes diferenças terminológicas entre as instituições, a quantidade de conceitos exclusivos entre elas é em média 6% dos conceitos.

Analisando a resposta dos avaliadores e, considerando-se as duas primeiras questões, fica evidente que a avaliação é baseada na experiência pessoal de cada um e não em um conhecimento formalizado, porém eles são unanimes no quesito das informações mínimas para o laudo radiológico, corroborando com o exposto por Prestes-Jr & et.al. em 2012: técnica do exame, descrição dos achados e impressão diagnóstica. O avaliador A10 foi bastante preciso em descrever a informação contida na categoria textual "Impressão Diagnóstica" ao escrever:

> *"...a avaliação global do exame disposta de forma mais concisa e objetiva dos achados, associada à minha interpretação pessoal no momento de responder aos questionamentos do medico solicitante." (A10)*

Em concordância com o que escreve Prestes-Jr:

> *"Nos casos em que o imaginologista está convicto de sua interpretação, o laudo pode ser concluído de forma mais direta e objetiva."* (Prestes-Júnior; Tourinho; Rangel, 2012).

Segundo as respostas dos avaliadores, pode se concluir que uma superestrutura de laudo radiológico desejável seria composta, de forma explicita, as partes constituintes "Técnica do Exame", "Descrição", "Impressão diagnóstica" e "Observação" quando necessário.

Três termos foram considerados inadequados pelos avaliadores em suas respectivas instituições, são eles: “espondilose”, “campo visual” e “hidrossiringomielia”. Dos três termos “espondilose” foi considerado inadequado por dois avaliadores. Para verificar e entender essas inadequações as referências destes termos na ferramenta *meta-genaralis* e um radiologista especialista foram consultados. O termo “espondilose” foi conceituado pelo DecS e o especialista explica que a confusão está no texto “qualquer parte das vértebras pode ser afetada” porque este termo é usado quando apenas a vértebra é afetada, portanto, neste caso, o DecS trás a descrição correta do conceito e não há equivoco na sua conceitualização. O termo composto “campo visual”, conhecido também por FOV do inglês “Field Of View”, foi conceituado utilizando o Dicionário Médico Ilustrado Dorland (Dorland, 1999) e, segundo especialista consultado, trata-se de uma confusão de tradução porque o termo se refere à imagem que está sendo estudada. O termo “hidrossiringomielia” foi conceituado por um artigo científico (Honorato; Borges, 2004) e, o especialista, assim como o avaliador A5, explica que o exame de ressonância magnética não permite distinguir entre hidromielia e siringomielia.

A metodologia *meta-generalis* é constituída por seis passos: 1) coleta dos laudos radiológicos de determinado tipo de exame; 2) decomposição textual; 3) normalização das unidades lexicais; 4) identificação das superestruturas textuais; 5) conceituação dos termos candidatos; 6) avaliação das superestruturas identificadas e da terminologia extraída pelos especialistas na linguagem de especialidade estudada.

Embora a metodologia tenha sido criada com base em uma amostra de exame e região anatômica específica, no caso, exame de ressonância magnética da coluna lombossacra, a ferramenta *meta-generalis* prevê a criação de contextos e idiomas diferentes, porque a ideia implícita do estudo é a de extrapolar a metodologia para

outros exames radiológicos e relatório médicos, avaliando a aplicabilidade e o aproveitamento resultante do produto terminológico eletrônico já construído por este estudo, e pelos estudos subsequentes, que poderão ser realizados.

Uma rede ontológica de decisão contextual foi construída pela aplicação da metodologia neste estudo prevendo a possibilidade de ser aplicada na recuperação de informação. Todavia, avaliações neste sentido não foram implantadas e outras pesquisas neste campo seriam de grande importância para verificar o quanto a decisão ontológica contextual, e a própria ontologia, melhora a recuperação das informações no corpus textual estudado. Além disso, exportação e importação da ontologia para ferramentas computacionais destinadas a criação e manutenção de ontologias, como por exemplo o Protegé, também seriam interessantes em pesquisas futuras porque poderiam propiciar a interoperabilidade entre diferentes sistemas (Deus; Cabral, 2011).

Seguindo a ideia desenvolvida por Freire em 2012, a estruturação de dados a partir de texto livre ou semiestruturado com explicitação do conhecimento lhe atribui semântica informacional, transformando o dado em informação computável e, promove o sistema de informação para o terceiro nível das etapas para geração do conhecimento e inteligência.

6 CONCLUSÃO

A metodologia utilizada permitiu a estruturação da informação considerando o idioma português e foi capaz de extrair as unidades lexicais mantendo as referencias com as suas respectivas partes constituintes e as conceituando ontologicamente. A superestrutura identificada vai ao encontro das necessidades dos avaliadores.

Portanto, é possível estruturar informações contidas em texto livre considerando sua superestrutura textual e a terminologia empregada em português.

7 BIBLIOGRAFIA

ACR. **ACR Breast Imaging and Radiological Data System Birads.** 3rd. ed. Reston, VA, USA.: [s.n.].

ACR, A. C. OF R. **ACR Practice Guideline for Communication of Diagnostic Imaging Findings.**, 2010. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:ACR+PRACTICE+GUIDELINE+FOR+COMMUNICATION+OF+DIAGNOSTIC+IMAGING+FINDINGS#9>. Acesso em: 24 jun. 2013

ADOBE, S. **Lorem and more.**, 2003. Disponível em: <http://www.adobe.com/cfusion/exchange/index.cfm?ev>

AGGARWAL, C. C.; CHENG XIANG ZHAI. An introduction to text mining. *In*: AGGARWAL, C. C.; ZHAI, C. (Eds.). **Mining Text Data**. Boston, MA: Springer US, 2012. p. 1–10.

ALMEIDA, G. M. DE B. FAZER TERMINOLOGIA É FAZER LINGUÍSTICA. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l.] EDIPUCRS, 2010. p. 72.

ALMEIDA, M. B.; SOUZA, R. R. Avaliação do espectro semântico de instrumentos para organização da informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 16, n. 31, 25 maio. 2011.

BARBOSA, F. *et al.* Radiological reports: a comparison between the transmission efficiency of information in free text and in structured reports. **Clinics (São Paulo, Brazil)**, v. 65, n. 1, p. 15–21, jan. 2010.

BELLATRECHE, L.; AIT-AMEUR, Y.; CHAKROUN, C. A design methodology of ontology based database applications. **Logic Journal of IGPL**, v. 19, n. 5, p. 648–665, 20 maio. 2010.

BIREME. **Guia de atualização do DeCS.** [S.l: s.n.]. p. 30

BORST, P.; AKKERMANS, H.; TOP, J. Engineering ontologies. **International Journal of Human-Computer Studies**, v. 46, n. 2-3, p. 365–406, fev. 1997.

BRAGA, T. M. **Uma ferramenta de mineração de texto em bancos de dados de um hospital universitário utilizando decomposições matriciais.** [S.l.] Universidade Federal do Rio de Janeiro, 2011.

BUBENHOFER, Noah; PTASHNYK, Stefaniya; KILIAN, Cristiane Krause. **Corpora, bancos de dados e a web: estado da arte da pesquisa em fraseologia e lexicografia assistida por computador**. Entrelinhas, v. 6, n. 1, p. 83-96, 2012.

CAVALINI, Luciana Tricai; COOK, Timothy Wayne. Engenharia do conhecimento de sistemas de informação em saúde com base em modelos multinível minimalistas. **Jornal Brasileiro de TeleSSaúde**, v. 1, n. 2, p. 43-50, 2012.

CAVALCANTE, R. B.; SILVA, P. C.; FERREIRA, M. N. Sistemas de informação em saúde: Possibilidades e Desafios. **Revista de Enfermagem da UFSM**, v. 1, n. 2, p. 290–299, 2011.

CEUSTERS, W.; SMITH, B. A unified framework for biomedical terminologies and ontologies. **Studies in health technology and informatics**, v. 160, n. Pt 2, p. 1050–1054, 2010.

CIMIANO, P.; REYLE, U.; ŠARIĆ, J. Ontology-driven discourse analysis for information extraction. **Data & Knowledge Engineering**, v. 55, n. 1, p. 59–83, out. 2005.

CHIBA, C.; NARDI, A. Desenvolvimento em camadas. **MSDN, 2007**. Disponível em: http://www.microsoft.com/brasil/msdn/tecnologias/arquitetura/Layers_Developing.mspx. Acesso em: 20 de jun. 2012.

CONRADO, M. DA S.; GUTIÉRREZ, V. A. L.; REZENDE, S. O. Evaluation of Normalization Techniques in Text Classification for Portuguese. *In*: MURGANTE, B. *et al.* (Eds.). **Computational Science and Its Applications – ICCSA 2012**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 7335p. 618–630.

CORRÊA, R.; MIRANDA, D. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **A.toZ: novas práticas em informação e conhecimento**, v. 1, n. 1, p. 11–22, 2011.

DBDESIGNER. **DBDesigner Fork**. Disponível em: <http:sourceforge.net/projects/dbdesigner-fork/>. Acesso em: 18 jun. 2012.

DENA, C.; DOUGLASS, J.; MARINO, M. **Benchmark fiction: A framework for comparative new media studies.**In Proceedings of the 6th Digital Arts and Culture Conference (DAC 2005). **Anais**...2005Disponível em: <http://www.christydena.com/wp-content/uploads/2007/11/2005-dac-benchmark-fiction-final-print-submission1.pdf>. Acesso em: 24 jun. 2013

DEUS, W. DE; CABRAL, E. Um estudo de ontologias como ferramenta para a modelagem do processo de comunicação entre agentes. **Revista Crase. edu**, v. 01, n. 1, p. 16–37, 2011.

DORLAND, W. A. N. **Dicionário Médico Ilustrado Dorland**. 28ª. ed. Barueri-SP: Editora Manole, 1999. p. 2217

DUQUE, J. L. **Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico.** [S.l: s.n.].

DUQUE, J.; MATOS, P.; CIFERRI, C. DE A. Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico. v. d, 2012.

ELDER, J.; MINER, G.; NISBERT, B. **Practical text mining and statistical analysis for non-structured text data applications.** Oxford: Elsevier Inc., 2012. p. 1053

FAULSTICH, E. Socioterminologia: mais que um método de pesquisa, uma disciplina. **Ciência da Informação**, v. 24, 1995.

FENELON, S. Aspectos ético-legais em Imaginologia. **Radiologia Brasileira**, v. 36, n. 1, 2003.

FRANCELIN, M. M.; KOBASHI, N. Y. Concepções sobre o conceito na organização da informação e do conhecimento. **Ciência da Informação**, v. 40, n. 2, p. 207–228, 2011.

FREIRE, P. DE S. *et al.* Memória organizacional e seu papel na gestão do conhecimento. **Ciências da Administração**, v. 14, n. 33, p. 41–51, 2012.

GALVÃO, M.C.B.; RICARTE, I.L.M. Prontuário do Paciente. 1ª ed. Rio de Janeiro-RJ: Editora Guanabara Koogan, 2012;

HONORATO, D.; BORGES, W. Syringohydromyelia associated to therapeutic procedures for severe forms of neurocysticercoses: case report. **Arq Neuropsiquiatr**, v. 62, n. 3-b, p. 885–888, 2004.

HUSSAIN, M. W.; AFZAL, M. T.; WAQAS, M. Finding Semantic Relationship among Associated Medical Terms. **Int. J. Emerg. Sci**, v. 2, n. 2, p. 300–309, 2012.

IHTSD, I. H. T. S. D. **SNOMED CT User Guide.** Disponível em: <www.snomed.org/ug.pdf>. Acesso em: 28 dez. 2012.

KAHN, C. E. Multilingual retrieval of radiology images. **Radiographics: a review publication of the Radiological Society of North America, Inc**, v. 29, n. 1, p. 23–9, 2009.

KONG, Z.; ZHANG, K. **Summarizing Keyword Search in Relational Database.**Third International Conference on Intelligent Human-Machine Systems and Cybernetics. **Anais**...IEEE, ago. 2011. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6038237>. Acesso em: 24 jun. 2013

KRAUSE, D. **Tópicos em Ontologia Analítica.** Departamento de filosofia da Universidade Federal de Santa Catarina. 2011. Disponível em: http://www.cfh.ufsc.br/~dkrause/pg/TopicosOntologia.pdf. Acesso em 14 jun. 2012.

LANGLOTZ, C. P. RadLex: a new method for indexing online educational materials. **Radiographics: a review publication of the Radiological Society of North America, Inc**, v. 26, n. 6, p. 1595–7, 2006.

MACHADO, A. DA S. *et al.* Reestruturação produtiva no setor de saúde: estudo de caso em um hospital de Fortaleza-CE. **Revista de Administração Hospitalar e Inovação em Saude**, p. 65–74, 2012.

MACIEL, A. M. B. Linguagens especializadas e terminologia: o passado projetando o futuro. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l.] EDIPUCRS, 2010. p. 28.

MACIEL-LIMA, S.; RASIA, J. As Imagens não Falam por si Mesmas: um estudo sociológico da equipe do transplante hepático Images. **Saúde Soc**, v. 21, n. 1, p. 42–51, 2012.

MARTINS, S. C.; MAURITTI, R.; COSTA, A. F. Acesso a Bases de Microdados: Aplicações e Impactos nas Pesquisas em Ciências Sociais. **Mediações-Revista de Ciências Sociais**, v. 18, n. 1, p. 66-82, 2013.

MARTINS, Valéria Farinazzo. **Avaliação de usabilidade para sistemas de transcrição automática de laudos em radiologia.** 2011. Tese (Doutorado em Sistemas Eletrônicos) - Escola Politécnica, University of São Paulo, São Paulo, 2011. Disponível em: <http://www.teses.usp.br/teses/disponiveis/3/3142/tde-20072011-092438/>. Acesso em: 2013-10-03.

MHAMDI, F.; ELLOUMI, M. A new survey on knowledge discovery and data mining. **2008 Second International Conference on Research Challenges in Information Science**, p. 427–432, jun. 2008.

MHIRI, S.; DESPRES, S. Ontology usability via a visualization tool for the semantic indexing of medical reports (DICOM SR). *In*: HOLZINGER, A. (Ed.). **HCI and Usability for Medicine and Health Care**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4799p. 409–414.

MICHAELIS. **MICHAELIS, Moderno Dicionário da Língua Portuguesa.** Disponível em: <http://michaelis.uol.com.br/moderno/portugues/index. php?>.

MICROSOFT DEVELOPER NETWORK (MSDN). Anatomy of a Unit Test. Disponível em: http://msdn.microsoft.com/en-us/library/ms182517. Acesso em 13 de mai. 2012.

NARDON, F. B. **Compartilhamento de conhecimento em saúde utilizando ontologias e bancos de dados dedutivos.** [S.l.] Escola Politécnica da Universidade de São Paulo, 2003.

NETTER, F. H. **Atlas de Anatomia Humana**. 2ª. ed. [S.l.] Artmed, 2000. p. 142

OLIVEIRA, A. DE; MOTTA, R. DA; CUNHA, G. Mineração de textos: uma experiência usando TMSK e RIKTEXT. p. 1–11, 2011.

OLIVEIRA, A. M. P. P. DE; ISQUERDO, A. N.; ALVES, I. M. **As ciências do léxico: lexicologia, lexicografia, terminologia.** [S.l.] Editora Ufms, 2007. v. 3

PADILHA, T. P. P.; LACERDA, A. N. **Text Recognition for Concept Maps Construction in Collaborative Environments.**Brazilian Symposium on Collaborative Systems. **Anais**...IEEE, out. 2012Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6391724>. Acesso em: 12 jun. 2013

PASSINI, M. L. C. **Mineração de textos para organização de documentos em centrais de atendimento.** [S.l: s.n.].

PASSMAN, L. J. *et al.* SISMAMA--implementation of an information system for breast cancer early detection programs in Brazil. **Breast (Edinburgh, Scotland)**, v. 20 Suppl 2, p. S35–9, abr. 2011.

PAVEL, S.; NOLET, D. **Manual de terminología.** [S.l.] Ministério de Obras Públicas e Governamentais do Canadá, 2002.

PINTO, A.; BRUNESE, L. Spectrum of diagnostic errors in radiology. **World journal of radiology**, v. 2, n. 10, p. 377–83, 28 out. 2010.

PINTO, T. L. **Aspectos da terminologia da economia internacional.** [S.l: s.n.].

PRESTES-JÚNIOR, L. C. L.; TOURINHO, E. K.; RANGEL, M. Análise médico-legal das demandas judiciais em imaginologia. **Radiologia Brasileira**, v. 45, n. 2, p. 98–100, 2012.

QUANTIN, C. *et al.* Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm. **Revue d'Épidémiologie et de Santé Publique**, v. 57, n. 1, p. e1–e6, fev. 2009.

RAHMAN, M. Information retrieval with text mining for Decision Support System. **Bangladesh Journal of Scientific Research**, v. 24, n. 2, p. 117–126, 2011.

REIS, A. DE O. As interfaces das disciplinas do Léxico. **Dialogia**, v. 9, n. 2, p. 173–179, 2010.

REZENDE, S.; MARCACINI, R.; MOURA, M. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, v. 7, p. 7–21, 2011.

ROBERTO, J.; MARTÍ, M.; LLORENTE, M. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. **Procesamiento de Lenguaje …**, v. 48, p. 97–104, 2012.

SCHAEFER, F. Proteção de dados de saúde como direito fundamental. **Cadernos da Escola de Direito e Relações Internacionais**, v. 17, p. 1–19, 2012.

SCHIESSL, Marcelo; BRÄSCHER, Marisa. Do texto às ontologias: uma perspectiva para a ciência da informação. **Ciência da Informação**, v. 40, n. 2, 2011.

SILVA-FILHO, S. C. DA. A VARIAÇÃO E A RELAÇÃO CONCEITO/TERMO: UMA QUESTÃO DE PONTO DE VISTA. **Trauma**, v. 6, n. 12, p. 75–86, 2010.

UMLS. **Unified Medical Language System**. Disponível em: <https://uts.nlm.nih.gov/home.html>. Acesso em: 19 jun. 2013.

VASILÉVSKI, V. Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português. **Linguamática**, v. 3, n. 2, p. 107–118, 2011.

VIEIRA, J. DE L.; CORREA, R. VISUALIZAÇÃO DA INFORMAÇÃO NA CONSTRUÇÃO DE INTERFACES AMIGÁVEIS PARA SISTEMAS DE

RECUPERAÇÃO DE INFORMAÇÃO. **Encontros Bibli**, v. 16, n. 32, p. 73–93, 2011.

VIEIRA, M. C. T. **Leitura no ensino superior: letramentos em pauta.**XV ENDIPE. **Anais**...2010

VIEIRA, R.; LOPES, L. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. *In*: **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l: s.n.]. .

VILLAVICÊNCIO, A.; RAMISCH, C. Chutando o balde ou batendo as botas? Processamento de linguagem natural e expressões multipalavra na linguagem científica. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. Porto Alegre: Edipucrs, 2010. p. 29–49.

ZERBINATTI, L. **Extração de Conhecimento em Laudos de Radiologia Torácica na Língua Portuguesa.** [S.l.] Escola Politécnica da Universidade de São Paulo, 2010.

ZILIO, L. TERMO E VALOR LINGUÍSTICO: UMA ABORDAGEM ENSAÍSTICA. **Cadernos do IL**, v. 42, p. 119–127, 2012.

8 APÊNDICES

8.1 APÊNDICE A: Termo de consentimento livre e esclarecido.

Você está sendo convidado(a) como voluntário(a) a participar de uma pesquisa chamada “Metodologia para estruturação de informações de laudos radiológicos”. Sua participação consiste no preenchimento de um questionário com 8 (oito) perguntas sobre informações extraídas de laudos radiológicos da coluna lombo sacra. Este estudo está sendo realizado por mim, Flavio Barbosa, aluno de doutorado do Departamento de Clínica Médica da Faculdade de Medicina de Ribeirão da USP.

Justificativa: A padronização diminui a variação das descrições dos achados no registro do laudo melhorando a transmissão das informações entre o médico que pede o exame e o que realiza consequentemente diminuindo a chance de erros na interpretação do exame.

Objetivo: Elaborar uma metodologia capaz de detectar e organizar a informação dos laudos radiológicos diminuindo a variação terminológica.

Procedimento para coleta de dados: A terminologia foi extraída de forma automatizada e semi-automatizada de laudos obtidos em três instituições/serviço de diagnostico por imagem diferentes no primeiro trimestre de 2010 e primeiro trimestre de 2011, portanto trata-se de dados retrospectivos que preservam a identidade dos pacientes e profissionais envolvidos.

Benefícios: A padronização dos termos melhora a transmissão das informações entre o médico solicitante e executante, minimizando as chances de erros no tratamento do paciente. Além disso, permitirá a recuperação dos laudos por termos específicos melhorando a pesquisa em saúde.

Risco: Não existem riscos para nenhum dos indivíduos.

Privacidade: As informações obtidas neste estudo poderão ser divulgadas em congressos e publicações científicas, sem nunca identificar os nomes dos participantes. Após a pesquisa os dados serão guardados por 5 (cinco) anos com o pesquisador e orientador da pesquisa Prof. Dr. Valdair Francisco Muglia.

Eu, Dr(a) _____, após a leitura deste documento e ter tido a oportunidade de conversar com o pesquisador responsável, para esclarecer todas as minhas dúvidas, acredito estar suficientemente informado, ficando claro para mim que minha participação é voluntária e que posso retirar este consentimento a qualquer momento sem penalidades ou perda de qualquer benefício. Estou ciente também dos objetivos da pesquisa e garantias de confidencialidade e esclarecimentos sempre que desejar mesmo que as respostas afetem minha vontade de continuar participando da pesquisa.

Diante do exposto expresso minha concordância de espontânea vontade em participar deste estudo:

_____
_____/____/_____.
Assinatura do Voluntário Data de Concordância

_____ _____
_____/____/_____.
Nome do Pesquisador Assinatura do Pesquisador Data da Assinatura

Contatos do pesquisador Flavio Barbosa: Telefone (16) 3602-2118 e e-mail: contato@flaviobarbosa.com.br.

8.2 APÊNDICE B: Questionário de avaliação da superestrutura e terminologia das informações radiológicas explicitadas.

CONSIDERE AS ESTRUTURAS DE LAUDO A SEGUIR:

MODELO A:
- Descrição
  - Observações (quando necessário)
- Conclusão

MODELO B:
- Descrição do laudo
  - Técnica do exame no primeiro paragrafo
- Impressão diagnóstica

MODELO C:
- Técnica do Exame
- Descrição
- Resultado

1. **Qual a estrutura de laudo mais adequada?**
   ( ) Modelo A
   ( ) Modelo B
   ( ) Modelo C

2. **Comente sobre a sua escolha do modelo. O que levou você a escolher este modelo?**

3. **Você sugeriria outra estrutura para o modelo de laudo da coluna lombo sacra?**
   Descreva abaixo. (Se deixar em branco entenderemos que o modelo escolhido é perfeito para elaboração do laudo radiológico da coluna lombossacra)

CONSIDERE OS TERMOS E CONCEITOS A SEGUIR:

ACETABULAR:
Alusivo ou pertencente a acetábulo. Que tem forma de copo de boca larga.

ARACNOIDITE:
Inflamação aguda ou crônica da membrana aracnóide das meninges, geralmente envolvendo a medula espinhal ou base do cérebro. Este termo geralmente se refere a um processo inflamatório persistente caracterizado por espessamento da membrana aracnóide e das adesões durais. Entre os estados associados estão: cirurgia prévia, infecções, trauma, hemorragia subaracnóide

e irritação química. Os sinais clínicos variam de acordo com o local da inflamação, mas incluem neuropatias cranianas, radiculopatias e mielopatias.

BORDA:
Extremidade limite de uma superfície.
Sinônimo(s): bordo, extremidade, magens, marginal, margear, margem, margens.

CAGE:
É uma estrutura rígida geralmente preenchida com osso, utilizada para estabilizar a coluna.

CAMPO VISUAL:
Área dentro da qual os estímulos produzirão a sensação de visão com o olho em uma posição direta à frente.
Sinônimo(s): fov.

ESPONDILOSE:
Doença degenerativa da coluna vertebral que pode afetar qualquer parte das vértebras, o disco intervertebral, e o tecido mole circunjacente.

ESTADIAMENTO DE NEOPLASIAS:
Métodos que tentam expressar em termos replicáveis a extensão de neoplasias no paciente.
Sinônimo(s): estadiamento, reestadiamento.

EXAME:
Inspeção do corpo de um doente como meio para diagnosticar uma doença ou anormalidade, qualificada de acordo com o método empregado.

Sinônimo(s): exame atual, estudo, exame, exame de imagem, laudo.

EXTRAVENTRICULAR:
Situado, ou ocorrendo, fora de um ventrículo.

FISSURA:
Sulco ou fenda mais ou menos profunda que delimita diferentes regiões de um órgão; cissura. Fenda, racha, abertura. Fenda ou fratura nos ossos.
Sinônimo(s): fissura, ruptura, sulco.

FRATURA TRAUMÁTICA:
Resulta da aplicação de uma força sobre o osso que seja maior que a resistência deste. Pode ocorrer no local de um impacto (fratura direta), num local afastado da zona de impacto (por exemplo, fratura da clavícula após queda sobre a mão - fratura indireta), ou por contracção muscular violenta (fratura por tração muscular).
Sinônimo(s): sobrecarga, trauma, traumático.

HEMATOPOESE:

Desenvolvimento e formação de vários tipos de células sangüíneas. a hematopoese pode ocorrer na medula óssea (medular) ou fora dela (hematopoese extramedular).

Sinônimo(s): hematopoiese, hematopoietico.

HEMIVÉRTEBRA:

A hemivértebra é o exemplo mais típico de defeito de formação. A forma como ela está relacionada com outras vértebras e o desvio produzido no alinhamento do eixo vertebral são fundamentais para a decisão sobre o tipo de tratamento necessário.

HIDROSSIRINGOMIELIA:

A hidrossiringomielia é definida como uma dilatação longitudinal do canal central da medula espinhal cujo conteúdo é semelhante ao líquido cefalorraquidiano e que freqüentemente produz comprometimento neurológico.

Sinônimo(s): hidrosseringomielia, seringohidromielia, seringo-hidromielia.

HIPOPLASIA:

Desenvolvimento incompleto ou subdesenvolvimento de um órgão ou tecido; ele é de grau menos severo do que aplasia.

INCIDENTE:

Acontecimento desagradá-vel, mas de menor importância, que aparece no decorrer de um procedimento paralelo.

Sinônimo(s): incidental, incidentalmente.

LÍQUIDO SINOVIAL:

Líquido claro e viscoso secretado pela membrana sinovial. Contém mucina, albumina, gordura e sais minerais, servindo para lubrificar as articulações.

Sinônimo(s): sinovia.

LIQUOR CEREBRO ESPINHAL:

O líquido contido dentro dos quatro ventrículos do cérebro, o espaço subaracnóideo e o canal central da medula espinhal; formado pelos plexos coróides e o parênquima cerebral, ele circula através dos ventrículos para dentro do espaço subaracnóideo e é absorvido para dentro do sistema venoso.

Sinônimo(s): LCE, LCR, liquor, liquorica, liquorico.

MEGAPÓFISE:

É o aumento do processo transverso da última vértebra lombar. Normalmente, o processo transverso da última vértebra lombar não alcança a asa do ilíaco, existindo um espaço livre. Caso exista uma megapófise, o processo transverso forma uma neoartrose com a asa do ilíaco. A sintomatologia é dor unilateral localizada ao nível da neoartrose. O tratamento para a megapófise é geralmente sintomático e em raros casos, há necessidade de um tratamento cirúrgico com osteotomia do processo transverso.

Sinônimo(s): mega apofise, mega-apofise.

METAL:

Cada uma de um grande grupo de substâncias simples que ostentam um brilho característico, "brilho metálico", lustroso ou opaco, quando polidas; são boas condutoras de eletricidade e de calor, podem fundir-se e comumente são maleáveis ou dúcteis.

Associações: metálico, parafuso.

NÓDULO:

Uma pequena saliência ou intumescimento que é sólida e pode ser detectada pelo tato.

Sinônimo(s): multinodular, nodular.

NÓDULO DE SCHMORL:

Um nódulo visto em radiografias da coluna vertebral, devido ao prolapso de um núcleo pulposo para dentro de uma vértebra adjacente.

Sinônimo(s): schimorl, schmorl, schmorll, schrmol, shmorl.

OSTEOARTRITE:

Doença articular degenerativa e progressiva que é a forma mais comum de artrite, especialmente em pessoas idosas. Acredita-se que a doença não resulta do processo de envelhecimento, mas de mudanças bioquímicas e estresses biomecânicos que afetam a cartilagem articular. Na literatura estrangeira, é freqüentemente chamada de osteoartrose deformante. Os termos Osteoartrose, Osteoartrite e Artrose definem uma mesma doença que chega a acometer até 30% da população adulta acima dos 50 anos.

Sinônimo(s): Artrose, discartrose, osteaortrose, osteartrose, osteoartrite, osteoartrose.

PROXIMAL: O mais perto do centro, da origem ou da cabeça.

Sinônimo(s): proximalmente.

QUIMIOTERAPIA: Tratamento de certas doenças por meio de compostos químicos que eletivamente atuam sobre certos organismos patogênicos ou sobre certos órgãos doentes; iatroquímica, Química médica.

RUPTURA:

Fratura ou rompimento traumático ou forçoso de um órgão ou outra parte macia do corpo.

Sinônimo(s): fissura.

RUPTURA ANULAR:

Quando a degeneração do núcleo pulposo estiver acompanhada da erosão do anel fibroso teremos a ruptura interna do disco.

SIRINGOMIELIA:

Cavidades longitudinais da medula espinhal, com maior freqüência na região cervical, podendo se estender para múltiplos níveis espinhais. As cavidades são alinhadas por tecido gliogênico denso e podem estar associadas com neoplasias da medula espinhal, lesões traumáticas da medula espinhal e malformações vasculares. A siringomielia é identificada clinicamente por dor e parestesia, atrofia muscular das mãos e analgesia com termoanestesia de mãos

e braços, porém com a sensação tátil preservada (dissociação sensorial). Espasticidade das extremidades inferiores e incontinência também podem se desenvolver.

TEXTURA:
A estrutura e organização de um tecido ou órgão.
Sinônimo(s): tissular

TRABECULAR:
Que diz respeito às trabéculas.

1. **Os termos apresentados são adequados para a descrição de achados?**
   ( ) Concordo plenamente
   ( ) Concordo parcialmente
   ( ) Discordo plenamente
   ( ) Discordo parcialmente
   ( ) Não posso avaliar

1. **Qual a quantidade de termos apresentados que é empregada na sua instituição para elaboração do laudo?**
   ( ) Nenhum dos termos apresentados
   ( ) 1 e 8 termos
   ( ) 9 e 15 termos
   ( ) 16 e 23 termos
   ( ) 24 e 30 termos

2. **Dos termos e conceitos apresentados, existe algum que não corresponde ao significado empregado na sua instituição (clínica)? Qual ou quais?**

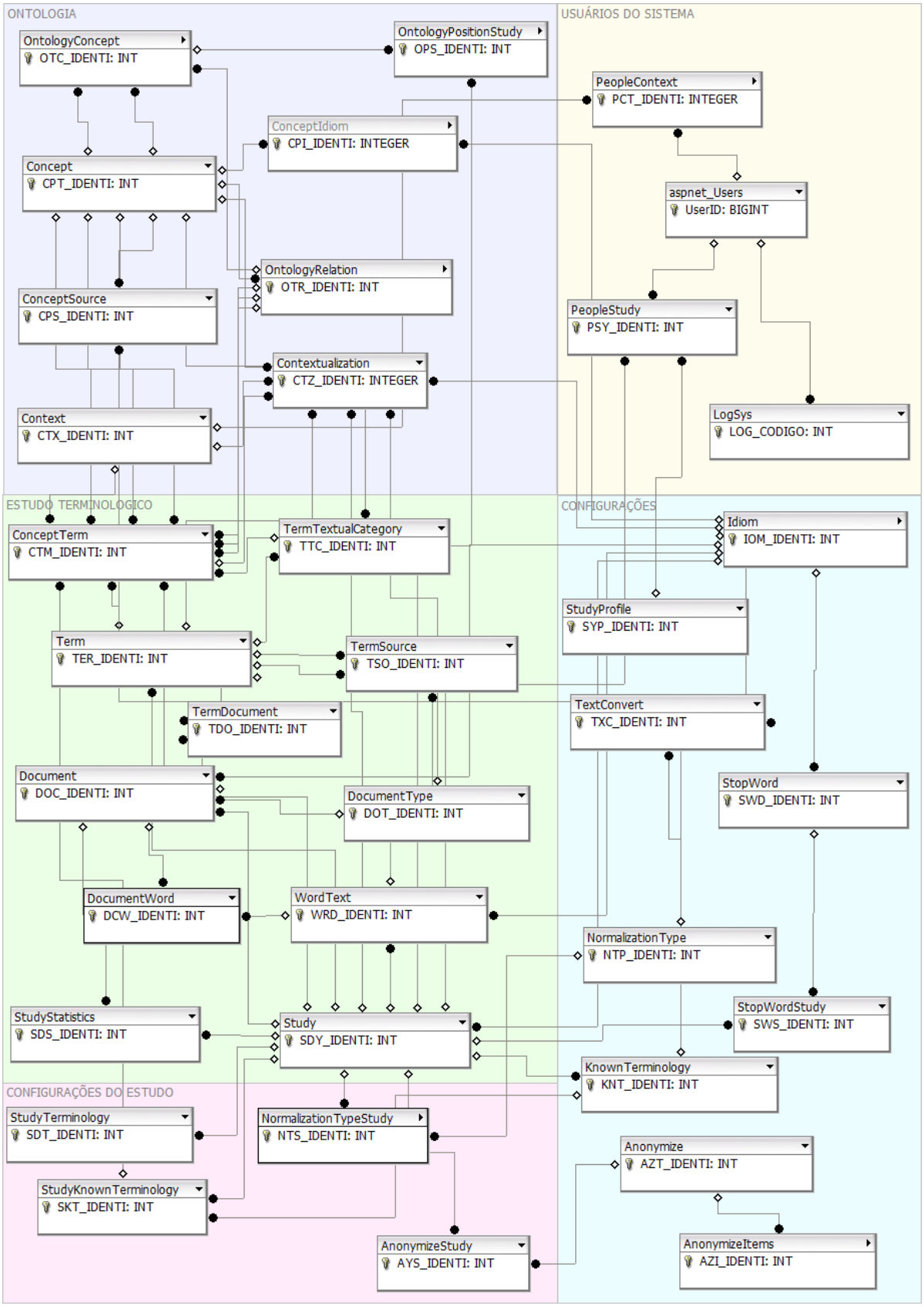
3. **Facilitaria a emissão do laudo radiológico a existência de um vocabulário comum entre diferentes instituições.**
   (Observe que se trata de uma afirmação)
   ( ) Concordo plenamente
   ( ) Concordo parcialmente
   ( ) Discordo plenamente
   ( ) Discordo parcialmente
   ( ) Não posso avaliar

4. **Por favor, se preciso, adicione seus comentários.**

8.3 APÊNDICE C: Distribuição dos termos a serem avaliados entre as instituições.

	Existe nas instituições:		
Termo	**A**	**B**	**C**
Acetabular	Sim	Não	Não
Aracnoidite	Sim	Sim	Não
Borda	Sim	Sim	Sim
Cage	Sim	Não	Não
Campo visual	Sim	Sim	Não
Espondilose	Sim	Sim	Sim
Estadiamento de Neoplasias	Sim	Não	Não
Exame	Sim	Sim	Sim
Extraventricular	Sim	Não	Não
Fissura	Sim	Sim	Sim
Fratura traumática	Sim	Sim	Não
Hematopoese	Sim	Sim	Não
Hemivértebra	Sim	Não	Não
Hidrossiringomielia	Sim	Não	Sim
Hipoplasia	Sim	Não	Sim
Incidente	Sim	Não	Sim
Líquido sinovial	Sim	Sim	Não
Liquor cerebro espinhal	Sim	Sim	Sim
Megapófise	Sim	Não	Sim
Metal	Sim	Sim	Sim
Nódulo	Sim	Sim	Sim
Nódulo de Schmorl	Sim	Sim	Sim
Osteoartrite	Sim	Sim	Sim
Proximal	Sim	Não	Sim
Quimioterapia	Sim	Não	Não
Ruptura	Sim	Sim	Sim
Ruptura anular	Sim	Não	Sim
Siringomielia	Sim	Não	Não
Textura	Sim	Sim	Não
Trabecular	Sim	Sim	Sim

8.4 APÊNDICE D: Modelo de entidade e relacionamento (MER) da ferramenta computacional *Meta-generalis*.

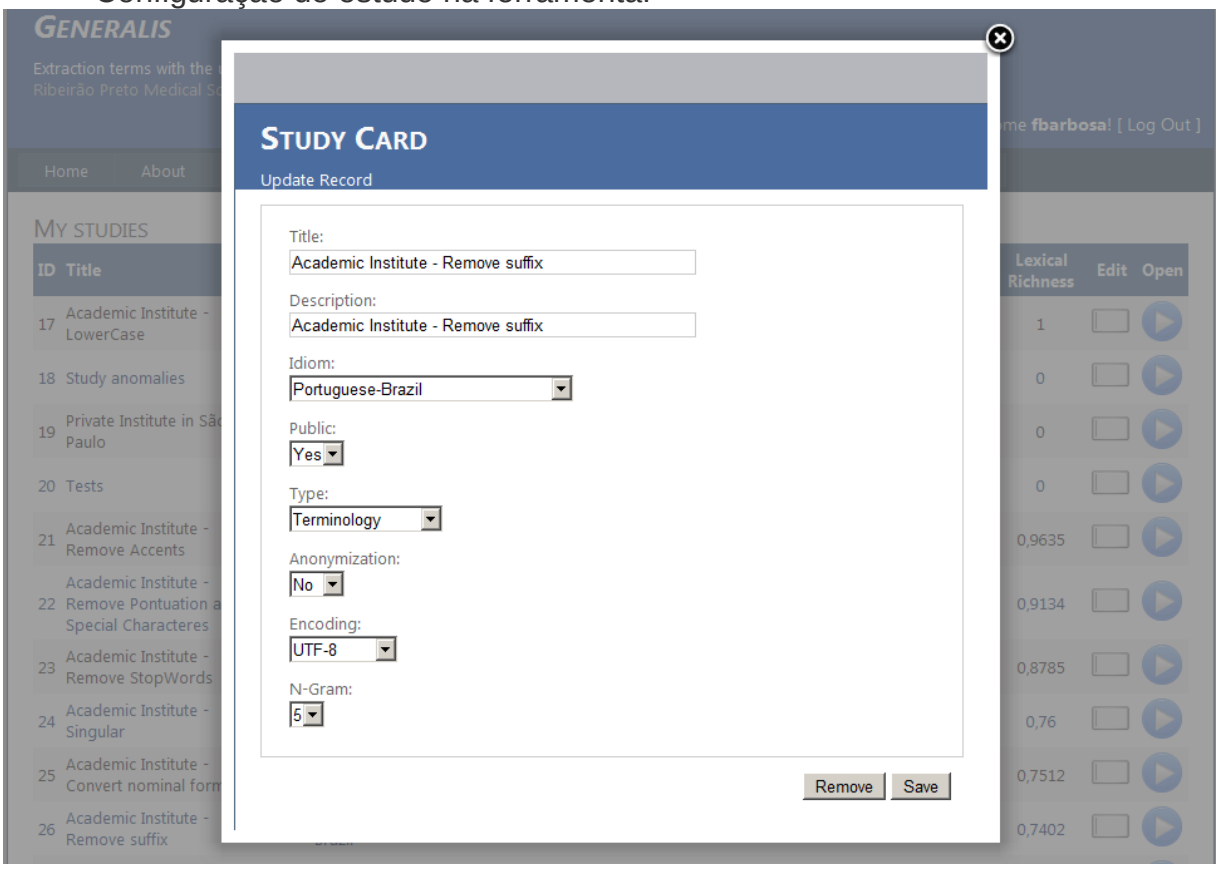


O MER foi divido nas seguintes áreas: ontologia, usuários do sistema, estudo terminológico, configurações e configurações do estudo. Foram criadas 34 tabelas, à exceção da tabela “aspnet_Users” própria do ASP.NET, conforme segue:

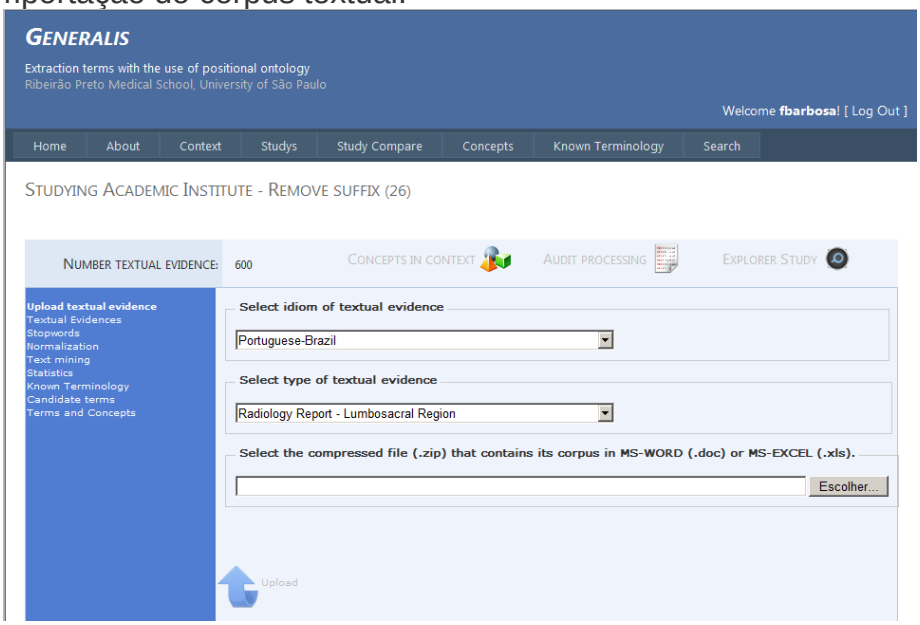
Tabela	**Responsabilidade**
Anonymize	Anonimizações disponíveis
AnonymizeItems	Itens de anonimização
AnonymizeStudy	Anonimizações configurados para o estudo
Aspnet_Users	Usuários controlados pelo ASP.NET
Concept	Conceitos
ConceptIdiom	Conceito em determinado idioma
ConceptSource	Origem do conceito
ConceptTerm	Termos associados ao conceito
Context	Contexto
Contextualization	Contextualização
Document	Documento
DocumentType	Tipos de documentos
DocumentWord	Origem das unidades lexicais por documento
Idiom	Idioma
KnownTerminology	Produtos terminológicos
LogSys	Auditorias do sistema
NormalizationType	Tipos de normalização
NormalizationTypeStudy	Tipos de normalização do estudo
OntologyConcept	Relações entre os conceitos
OntologyPositionStudy	Decisão ontológica contextual horizontal e vertical
OntologyRelation	Relações ontológicas
PeopleContext	Relação do usuário com o contexto
StopWord	Lista de stopwords
StopWordStudy	Lista de stopwords do estudo
Study	Estudo
StudyKnownTerminology	Produtos terminológicos utilizados no estudo para conceituação automática
StudyProfile	Perfis de acesso ao estudo
StudyStatistics	Estatísticas do estudo
StudyTerminology	Configurações adicionais para estudo terminológico
Term	Termos
TermDocument	Origem dos termos por documento
TermSource	Unidades lexicais que originaram o termo
TermTextualCategory	Agrupamento dos termos por categoria textual
TextConvert	Tipos de conversão textual utilizada pelos diferentes tipos de normalização
WordText	Unidades lexicais extraídas do corpus textual (documentos)

8.5 APÊNDICE E: Interfaces da ferramenta computacional *Meta-generalis*.

Configuração do estudo na ferramenta:



Importação do corpus textual:



Normalizações configuradas no estudo:

STUDYING ACADEMIC INSTITUTE - REMOVE SUFFIX (26)

NUMBER TEXTUAL EVIDENCE: 600 CONCEPTS IN CONTEXT AUDIT PROCESSING EXPLORER STUDY

Upload textual evidence
Textual Evidences
Stopwords
Normalization
Text mining
Statistics
Known Terminology
Candidate terms
Terms and Concepts

Normalization unit terminology

Normalization type: Remove accent before generating the terminological unit

Apply: Unit terminology

N-Gram (1-5): 1

Insert Copy from...

Normalization type	Apply	N-Gram	Date	Remove	Edit	Up	Down
Insert space in the terminological units ("words")	Pré-Processamento	1	08/05/2012 07:54:48				
Transform all terminological units lowercased	Palavras	1	08/05/2012 07:54:48				
Remove accent before generating the terminological unit	Palavras	1	08/05/2012 07:54:48				
Remove special characters and punctuation	Palavras	1	08/05/2012 07:54:48				
Remove stopwords before generating the terminology unit	Palavras	1	08/05/2012 07:54:48				
Transform plural to singular	Palavras	1	08/05/2012 07:54:48				
Transforms nominal forms to verbs	Palavras	1	08/05/2012 07:54:48				
Transforms or removes known suffix	Palavras	1	08/05/2012 07:55:14				
Transform all terminological units lowercased	Palavras	2	08/05/2012 07:54:48				
Remove accent before generating the terminological unit	Palavras	2	08/05/2012 07:54:48				

1 2 3 4

Lista de unidades lexicais normalizadas:

STUDYING ACADEMIC INSTITUTE - REMOVE SUFFIX (26)

NUMBER TEXTUAL EVIDENCE: 600 CONCEPTS IN CONTEXT AUDIT PROCESSING EXPLORER STUDY

Upload textual evidence
Textual Evidences
Stopwords
Normalization
Text mining
Statistics
Known Terminology
Candidate terms
Terms and Concepts

Source term: N-Gram: 1 Term type: All Textual Category: All
Concept: All Relation: All Ignore Textual Category: No
Frequency of the terms between: 1 and 99999 Frequency of the terms in the documents: 1 and 99999
Ordenation: Number of documents with the term Descendant Number Terms: 1949 Apply Filter
Export to Excel Set category textual Conceptualize Automatically

N-Gram	Termo	Repetition maximum in a document	Number of documents with the term	Frequency	It´s an index term	Associated with a concept	Concepts
1	conclusao	3	600	602	✓	✓	(is) Conclusão radiológica
1	descricao	2	600	602	✓	✓	(is) Descrição radiológica
1	sinal	8	590	1946	✓	✓	(part of) Descrição radiológica , (part of) Conclusão radiológica , (part of) Tecnica do Exame , (part of) Imagem por Ressonância Magnética , (part of) Patient , (is) sinal if before for proporção , (is) sinal if before for alteração , (is) sinal if after for normal , (is) hipersinal if before for aumento , (is) hipersinal if after for aumento , (part of) Observações e comentários , (is) hipossinal if before for discreto , (is) hipossinal if before for redução , (is) indicar
1	vertebral	14	590	1963	✓	✓	(part of) Descrição radiológica , (part of) Conclusão radiológica , (part of) Tecnica do Exame , (part of) Imagem por Ressonância Magnética , (part of) Patient , (part of) coluna vertebral , (part of) Região Anatômica , (part of) Observações e comentários

Conceitualização da unidade lexical, associação da unidade lexical ao conceito:

CONCEPTUALIZE TERM

Estudo	Academic Institute - Remove suffix
N-Gram	1
Identificador	1759442
Termo	conclusao
Frequência	602
Freq. em Documento	600
Source	CONCLUSÃO
Termo de Índice	Sim
Normalização	TRANSFORM.LOWERCASE, REMOVE.ACCENT, REMOVE.CHARACTER.SPECIAL.PUNCTUATION

Concepts in context Search the EXACT DECS in P Search Import

Associated concepts

Compatibility	Relation	Main Term	Concept definition	Context	Before Relation	Before Concept	After Relation	After Concept	Inicial	Final	ID Concept	Position	Disassociate	Edit
Compatible	IS	Conclusão radiológica	Informações sobre o diagnóstico dos pacientes com base nas imagens analisadas.						0	0	3			

Documents

Count: **600**

ID.	File	Description
20953	0006-0093.txt	DESCRIÇÃO:-Exame da coluna lombar realizado em equipamento de 1,5 Tesla, antes e após a administração endovenosa de contraste paramagnético: Retificação da lordose lombar. Discreta retrolistese de L2 sobre L3. Osteófitos marginais nos corpos vertebrais lombares e nódulo de Schmorl no platô inferior de L2. Imagens nodulares com hiperssinal nas imagens ponderadas em T1 e T2 no corpo vertebral de L2 e L3, que pode corresponder a hemangioma ou substituição gordurosa. Alteração de sinal junto às placas terminais de L5-S1, de aspecto degenerativo. Heterogeneidade da intensidade de sinal da medula óssea das vértebras estudadas. Perda do sinal habitual dos discos intervertebrais nas imagens ponderadas em T2. Redução da altura dos espaços discais intervertebrais L2-L3 e L5-S1. Abaulamento dos discos intervertebrais L1-L2, L2-L3 e L5-S1. Alterações degenerativas nas articulações facetárias. Proeminência dos ligamentos amarelos em L2-L3 a L5-S1. As alterações degenerativas descritas causam estreitamento dos forames intervertebrais de L5-S1. Cone medular e cauda eqüina com intensidade de sinal e volume dentro da normalidade. Lipossubstituição discreta da musculatura paravertebral posterior. - Realizado exame complementar da coluna total com aquisições panorâmicas sagitais: Osteófitos marginais nos corpos vertebrais cervicais. Alteração de sinal nos corpos vertebrais de C2, D2 e D3, com hipersinal nas ponderações T1 e T2, que podem corresponder a hemangioma ou substituição gordurosa. Perda do sinal habitual dos discos intervertebrais cervicais. Pequenos complexos disco-osteofitários de C3-4 a C5-6, tocando o saco dural, sem comprometimento foraminal. Alterações degenerativas das articulações facetárias. Intensidade de sinal normal da medula espinhal. **CONCLUSÃO**: 1.Alteração de sinal da medula óssea das vértebras estudadas, de aspecto inespecífico (que pode corresponder a suspeita clínica de mieloma). 2.Alterações degenerativas da coluna vertebral lombossacra com estreitamento dos forames intervertebrais de L5-S1. Sem alterações significativas em relação ao exame de coluna lombar de 18/02/2010. **CONCLUSÃO**:Vide **conclusão** acima. Sem alterações significativas em relação ao exame de coluna lombar de 18/02/2010.
20861	0006-0001.txt	DESCRIÇÃO:Fratura do corpo vertebral de L1, acometendo os pilares anterior, médio e posterior, com redução de altura e retropulsão de fragmento ósseo, causando redução do canal vertebral, principalmente na região central e paramediana esquerda e comprimindo a porção terminal do cone medular e início da cauda equina. Discreto borramento e alteração de sinal da medula espinhal distal, no nível de T12. Demais estruturas ósseas íntegras. Ausência de hérnias ou protrusões discais. Intensidade normal das facetas articulares. Musculatura paravertebral sem alterações.**CONCLUSÃO**:Fratura de L1 com redução do canal vertebral e compressão da cauda equina.
20862	0006-0002.txt	DESCRIÇÃO:Osteófitos marginais nos corpos vertebrais lombares. Perda do sinal habitual dos discos intervertebrais lombares nas imagens ponderadas em T2. Protrusão discal posterior difusa L2-L3 e L3-L4, em contato com o saco dural, reduzindo discretamente a porção inferior dos forames de conjugação. Protrusão discal posterior difusa L4-L5 associado a hipertrofia facetária e do ligamento amarelo, causando redução das dimensões do canal e forames de conjugação nesse nível. Pequena protrusão discal posterior central L5-S1. Alterações degenerativas nas articulações facetarias lombares. Notam-se também alterações degenerativas junto aos processos espinhosos intervertebrais. Cone medular e cauda eqüina com intensidade de sinal e volume dentro da normalidade. Musculatura paravertebral parcialmente lipossubstituída.**CONCLUSÃO**:1.Alterações degenerativas da coluna vertebral lombossacra com protrusões discais difusas em vários níveis. 2. Redução das dimensões do canal vertebral lombar e forames de conjugação em L4-L5, de aspecto degenerativo.

8.6 APÊNDICE F: Tabelas de normalização textuais.

Nome da Normalização	Nome da conversão textual	Configuração da normalização		
		Procurar por	Condição	Substituir por
Remoção de acentos	Troca letra com acento por letra sem acento			
Remoção de *stopwords*	Não incluí as *stopwords* na lista de termos candidatos (vide apêndice G)			
Remoção de caracteres especiais e de pontuação	Remove pontuação e caracteres especiais concatenados a unidade lexical.			
Transformação para minúsculo	Troca letra maiúscula por sua correspondente minúscula.			
Transformação das formas nominas em verbos	Gerúndio	"ndo"	Final da unidade lexical	"r"
Transformação das formas nominas em verbos	Participio	"ado"	Final da unidade lexical	"ar"
Transformação de sufixos conhecidos	Do sufixo "IVA" para "IVO"	"iva"	Final da unidade lexical	"ivo"
Transformação de sufixos conhecidos	Do sufixo "ICA" para "ICO"	"ica"	Final da unidade lexical	"ico"
Transformação do plural para o singular	De "oes" para "ao"	"oes"	Final da unidade lexical	"ao"
Transformação do plural para o singular	De "ais" para "al"	"ais"	Final da unidade lexical	"al"
Transformação do plural para o singular	De "as" para "a"	"as"	Final da unidade lexical	"a"
Transformação do plural para o singular	De "res" para "r"	"res"	Final da unidade lexical	"r"
Transformação do plural para o singular	De "es" para "e"	"es"	Final da unidade lexical	"e"
Transformação do plural para o singular	De "os" para "o"	"os"	Final da unidade lexical	"o"
Pré-processamento	Separando o ponto final	"."	Final da unidade lexical	" . "
Pré-processamento	Separando a unidade de centímetro	"cm"	Final da unidade lexical	" cm"

Pré-processamento	Separando a unidade de milímetro	"mm"	Final da unidade lexical	" mm"
Pré-processamento	Separando o sinal "maior que"	">"	Contém	" > "
Pré-processamento	Separando o sinal "menor que"	"<"	Contém	" < "
Pré-processamento	Substituindo tabulação por espaço	"\t"	Contém	" "
Pré-processamento	Substituindo quebra de linha por espaço	"\n"	Contém	" "
Pré-processamento	Adicionando espaço após abreviação DRA.	"DRA."	Contém	"DRA. "
Pré-processamento	Adicionando espaço após abreviação DR.	"DR."	Contém	"DR. "
Pré-processamento	Adicionando espaço após "ÇÃO:"	"ÇÃO:"	Contém	"ÇÃO: "
Pré-processamento	Adicionando espaço após "SÃO:"	"SÃO:"	Contém	"SÃO: "
Pré-processamento	Adicionando espaço após "CONCLUSÃO:"	"CONCLUSÃO:"	Contém	" CONCLUSÃO: "
Pré-processamento	Separando abertura de parentese	"("	Contém	" ("
Pré-processamento	Separando o fechamento de parentese	")"	Contém	") "
Pré-processamento	Separando abertura de chave	"["	Contém	" ["
Pré-processamento	Separando fechamento de chave	"]"	Contém	"] "
Pré-processamento	Separando abertura de colchete	"{"	Contém	" { "
Pré-processamento	Separando fechamento de colchete	"}"	Contém	" } "
Pré-processamento	Separando ":" da unidade lexical anterior e posterior	":"	Contém	" : "
Pré-processamento	Separando "?" da unidade lexical anterior e posterior	"?"	Contém	" ? "
Pré-processamento	Separando "!" da unidade lexical anterior e posterior	"!"	Contém	" ! "
Pré-processamento	Separando "#" da unidade lexical anterior e posterior	"#"	Contém	" # "

Pré-processamento	Separando "$" da unidade lexical anterior e posterior	"$"	Contém	" $ "
Pré-processamento	Separando "*" da unidade lexical anterior e posterior	"*"	Contém	" * "
Pré-processamento	Separando "&" da unidade lexical anterior e posterior	"&"	Contém	" & "
Pré-processamento	Separando "." da unidade lexical anterior e posterior	". "	Contém	" . "
Pré-processamento	Separando "," da unidade lexical anterior e posterior	", "	Contém	" , "
Pré-processamento	Separando "%" da unidade lexical anterior e posterior	"% "	Contém	" % "
Pré-processamento	Separando hífen	"-"	Final da unidade lexical	" -"
Pré-processamento	Separando marcador com hífen	"1-"	Começo da unidade lexical	"1- "
Pré-processamento	Separando marcador com hífen	"2-"	Começo da unidade lexical	"2- "
Pré-processamento	Separando marcador com hífen	"3-"	Começo da unidade lexical	"3- "
Pré-processamento	Separando marcador com hífen	"4-"	Começo da unidade lexical	"4- "
Pré-processamento	Separando marcador com hífen	"5-"	Começo da unidade lexical	"5- "
Pré-processamento	Separando marcador com hífen	"6-"	Começo da unidade lexical	"6- "
Pré-processamento	Separando marcador com hífen	"7-"	Começo da unidade lexical	"7- "
Pré-processamento	Separando marcador com hífen	"8-"	Começo da unidade lexical	"8- "
Pré-processamento	Separando marcador com hífen	"9-"	Começo da unidade lexical	"9- "
Pré-processamento	Separando marcador com ponto final	"1."	Começo da unidade lexical	"1. "
Pré-processamento	Separando marcador com ponto final	"2."	Começo da unidade lexical	"2. "
Pré-processamento	Separando marcador com ponto final	"3."	Começo da unidade lexical	"3. "
Pré-processamento	Separando marcador com ponto final	"4."	Começo da unidade lexical	"4. "

Pré-processamento	Separando marcador com ponto final	"5."	Começo da unidade lexical	"5. "
Pré-processamento	Separando marcador com ponto final	"6."	Começo da unidade lexical	"6. "
Pré-processamento	Separando marcador com ponto final	"7."	Começo da unidade lexical	"7. "
Pré-processamento	Separando marcador com ponto final	"8."	Começo da unidade lexical	"8. "
Pré-processamento	Separando marcador com ponto final	"9."	Começo da unidade lexical	"9. "
Pré-processamento	Separando símbolo ordinal	"º"	Final da unidade lexical	" º"

8.7 APÊNDICE G: Lista de unidades lexicais ignoradas (stopword) em português.

a, acaso, acerca, acola, ademais, afinal, afora, agora, agorinha, ai, ainda, alem, algo, alguem, ali, alias, amiude, ao, aos, apud, aquela, aquelas, aquele, aqueles, aquilo, as, assim, cade, cadern, comigo, conosco, consigo, contigo, cuja, cujas, cujo, cujos, da, dai, dali, dantes, daquela, daquelas, daquele, daqueles, daquilo, das, de, dela, delas, dele, deles, demais, dessa, dessas, desse, desses, desta, destas, deste, destes, deveras, disso, disto, do, donde, doravante, dos, dum, duma, dumas, duns, durante, e, eis, ela, elas, ele, eles, em, embora, enfim, enquanto, entanto, entao, entretanto, essa, essas, esse, esses, esta, estas, este, estes, eu, felizmente, foi, foram, gracas, hoje, ibidem, idem, inda, infelizmente, isso, isto, ja, jamais, la, lha, lhas, lhe, lhos, logo, mais, malgrado, mas, me, mesma, mesmas, mesmo, mesmos, meu, meus, mim, minha, minhas, nalgum, nalguma, nalgumas, nalguns, naquela, naquelas, naquele, naqueles, naquilo, nas, nela, nelas, nele, neles, nem, nessa, nessas,

nesse, nesses, nesta, nestas, neste, nestes, ninguem, nisso, nisto, nos, nossa, nossas, nosso, nossos, noutra, noutras, noutro, noutros, novamente, num, numa, numas, nunquinha, nuns, o, ontem, ora, os, ou, outra, outras, outrem, outro, outrora, outros, outrossim, para, pela, pelas, pelo, pelos, per, pero, pois, por, porem, porquanto, porque, portanto, porventura, possivelmente, posteriormente, posto, pra, praquela, praquelas, praquele, praqueles, praquilo, pras, praticamente, prela, prelas, prele, preles, preste, prestes, propria, proprias, proprio, proximo, quais, quaisquer, qual, qualquer, quando, quanta, quantas, quanto, quantos, quao, que, quem, quer, quiça, realmente, recentemente, salvante, salvo, se, segundo, seja, sempre, senao, sequer, seu, seus, sim, simplesmente, sobremaneira, sobremodo, sobretudo, somente, sua, suas, tais, tal, tambem, tampouco, tao-so, tao-somente, te, teu, teus, ti, tirante, toda, todas, todavia, todo, todos, tu, tua, tuas, tudo, voce, voces, vos, vossa, vossos, vulgo.

9 ANEXOS

9.1 ANEXO A: Autorização de acesso aos dados da amostra.

HC 55
FMRP-USP
HOSPITAL DAS CLÍNICAS DA
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO - USP

AUTORIZAÇÃO

Declaremos para os devidos fins que o discente Flávio Barbosa, RG 34.770.206-5 SSP/SP, doutorando do curso de pós-graduação em Ciências Médicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, dispões de autorização e consentimento para acessar os laudos radiológicos do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto. O mesmo consentimento é dado para que utilizem as informações coletadas por meio da leitura e análise dos laudos para pesquisa de trabalho de doutorado sobre a temática "Metodologia para Estruturação de Laudos Radiológicos". Ressaltamos, no entanto, a necessidade de preservar as identidades de nossos clientes.

Ribeirão Preto, 01 de fevereiro de 2011.

Responsável
Prof. Dr. Jorge Elias Júnior
Coordenador do Centro de Ciências das Imagens e Física Médica

cedirp
TECNOLOGIA A SERVIÇO DA VIDA

Central de Diagnóstico
Ribeirão Preto

AUTORIZAÇÃO

Declaremos para os devidos fins que o discente Flávio Barbosa, RG 34.770.206-5 SSP/SP, doutorando do curso de pós-graduação em Ciências Médicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, dispões de autorização e consentimento para acessar os laudos radiológicos da CEDIRP. O mesmo consentimento é dado para que utilizem as informações coletadas por meio da leitura e análise dos laudos para pesquisa de trabalho de doutorado sobre a temática "Metodologia para Estruturação de Laudos Radiológicos". Ressaltamos, no entanto, a absoluta necessidade de sigilo total para preservação das identidades de nossos clientes.

Ribeirão Preto, 17 de novembro de 2011.

Eduardo Simão Trad

Omnimagem
diagnóstico para uma vida saudável

AUTORIZAÇÃO

Declaremos para os devidos fins que o discente Flávio Barbosa, RG 34.770.206-5 SSP/SP, doutorando do curso de pós-graduação em Ciências Médicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, dispõe de autorização e consentimento para acessar os laudos radiológicos da Omnimagem. O mesmo consentimento é dado para que utilizem as informações coletadas por meio da leitura e análise dos laudos para pesquisa de trabalho de doutorado sobre a temática "Metodologia para Estruturação de Laudos Radiológicos". Ressaltamos, no entanto, a necessidade de preservar as identidades de nossos clientes.

Fortaleza, 01 de fevereiro de 2011.

José Daniel Vieira de Castro

CREMEC 6950

Diretor-técnico

9.2 ANEXO B: "Generalis: extração e conceitualização de termos apoiado por terminologias e ontologia posicional" (Artigo publicado e apresentado oralmente durante o CBIS2012)

III Congresso Brasileiro em Informática em Saúde – CBIS 2012

GENERALIS: EXTRAÇÃO E CONCEITUALIZAÇÃO DE TERMOS APOIADO POR TERMINOLOGIAS E ONTOLOGIA POSICIONAL

Flávio Barbosa[1] e Valdair Francisco Muglia[2]

[1] Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP), Brasil

Resumo: O nome "*generalis*" é uma referência a um conceito da filosofia denominado *metaphysica generalis* que se ocupa da natureza e constituição da realidade e que foi chamada mais tarde de ontologia. No contexto de tecnologia da informação, vários artigos – Cimiano P & et.al., 2005; Mhiri S & Despres S, 2007; Bellatreche L & et.al., 2010 - definem ontologia como uma especificação explícita de uma conceitualização. *Generalis* é uma plataforma em ambiente WEB que permite a colaboração entre usuários de um determinado produto terminológico (estudo) e alimenta uma base única de conhecimento baseado em ontologia posicional. Ela é capaz de decompor documentos contendo texto não-estruturado em unidades lexicais (tokens) utilizando técnicas de mineração de texto e os reconstruir em uma estrutura informacional conceitualizada utilizando ontológica posicional e apoiada por produtos terminológicos como o DeCS (descritores em ciências da saúde).

Palavras-chave: Ontologia posicional, Terminologia, Mineração de Texto, Laudo Radiológico e Texto Livre.

***Abstract**: The name "generalis" is a reference to a concept known as "metaphysica generalis" of philosophy that deals with the nature and constitution of reality and what was later called ontology. In the context of information technology, many articles - Cimiano P & et.al., 2005; Mhiri Despres S & S, 2007, L & Bellatreche et.al., 2010 - defines ontology as an explicit specification of a conceptualization. Generalis is a platform in a web environment that enables collaboration between users of a terminology product (study) and maintains a unique base of knowledge based on positional ontology. This platform is able to decompose documents containing unstructured text in lexical units using text mining techniques and reconstructs them in an informational structure using positional ontological conceptualized and supported by terminological products such as DeCS (Descriptors in Health Sciences).*

***Keywords**: Positional ontology, Terminology, Text mining, Radiology report and Freetext.*

Introdução

O nome "*generalis*" é uma referência a um conceito da filosofia denominado metaphysica *generalis* que se ocupa da natureza e constituição da realidade e que foi chamada mais tarde de ontologia [1]. No contexto de tecnologia da informação, vários artigos – Cimiano P & et.al., 2005 [2]; Mhiri S & Despres S, 2007 [3]; Bellatreche L & et.al., 2010 [4] - definem ontologia como uma especificação explícita de uma conceitualização [5].

Existem duas definições para terminologia: 1) enquanto atividade é o conjunto vocabular, conjunto de léxicos, próprio de uma ciência; 2) enquanto teoria é o conjunto de postulados teóricos necessários para dar suporte à análise de fenômenos linguísticos concernentes à comunicação especializada. Para uma comunicação especializada é necessário à utilização de uma linguagem especializada que ocorre quando um conteúdo é descrito de maneira precisa, com referência a um esquema científico determinado ou uma estrutura normativa estabelecida, o conteúdo torna-se objeto de conhecimento especializado [6].

Laudos radiológicos são documentos produzidos por comunidade profissional especializada (médicos radiologistas) que descrevem de maneira precisa os achados observados nas imagens médicas de acordo com o conhecimento existente em referencial teórico cientifico. Portanto, são documentos que utilizam linguagem especializada e são constituídos por unidades lexicais (tokens) que referenciam conceitos da atividade médica radiológica. Ora, termos são unidades lexicais que representam conceitos [6], portanto os laudos radiológicos são documentos constituídos de termos.

As unidades lexicais são constituídas por uma ou mais palavras e, portanto, possuem uma espécie de gramatura. Quando uma unidade lexical é constituída por apenas uma palavra a gramatura é "1" e a notação é escrita da seguinte forma "nGram=1". Se uma unidade lexical é constituída de duas palavras a gramatura é "2" e a notação é escrita da seguinte forma "nGram=2" e recebe a denominação de bigrama. Quando a unidade lexical é constituída por três palavras a gramatura é "3" e a notação é "nGram=3" e recebe a denominação de trigrama e assim sucessivamente [6].

O objetivo deste trabalho é apresentar a plataforma *generalis* e observar a viabilidade de construir softwares para mineração de texto e construção de ontologias utilizando banco de dados relacionais (RDB – Relational database) [7].

Métodos

Os documentos utilizados são laudos radiológicos da coluna lombo sacra elaborados consecutivamente no primeiro trimestre de 2010 e primeiro trimestre de 2011. Para formar o corpus textual a ser analisado, foram coletados 600 laudos radiológicos de uma instituição privada e outros 600 de uma instituição pública acadêmica totalizando 1.200 laudos. Nestes laudos não havia qualquer informação que pudesse identificar o paciente, médico executante ou médico solicitante e, portanto, são anônimos. O trabalho foi submetido ao comitê de ética em pesquisa clínica e recebeu o número CAAE 00593812.2.1001.5440.

As seguintes normalizações lexicais foram aplicadas com a finalidade de gerar os termos candidatos (unidades lexicais normalizadas): 1) conversão para minúsculo; 2) remoção de acentos; 3) remoção de caracteres especiais e de pontuação; 4) remoção de stopword; 5) conversão de plural para singular; 6) conversão de formas nominais para verbos; 7) conversão e remoção de sufixos. Para cada tipo de normalização lexical foi criado um estudo específico dentro da plataforma *generalis* com o objetivo de observar cada método de normalização das unidades lexicais. Embora a plataforma *generalis* permita a geração de termos candidatos com gramatura igual a cinco (nGram=5), neste estudo as observações são referentes a gramatura "1" (nGram=1).

No momento da análise dos termos candidatos para conceitualização foi solicitada à plataforma *generalis* a geração automática dos conceitos utilizando um produto terminológico conhecido: o DeCS (descritores em ciências da saúde) [8]. Para este estudo duas consultas foram feitas ao DeCS usando o idioma português: uma por unidade lexical exata e outra por unidade lexical semelhante [9].

A arquitetura de software empregada para plataforma *generalis* é de camadas [10], como segue: acesso a dados, regra de negócios, testes unitários [11] e interface. As camadas de acesso a dados, regra de negócios e testes unitários ficam dentro de um único projeto do tipo biblioteca de classe (class library). A figura 1 ilustra a relação entre o banco de dados relacional e as camadas da plataforma.

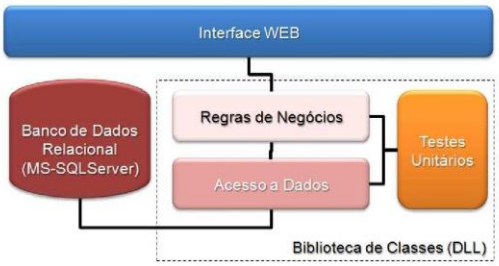


Figura 1: Arquitetura de software da plataforma *generalis*

Com a exceção da ferramenta DBDesigner Fork 2009 [12] para desenho do modelo de entidade e relacionamento (MER – Model Entity Relation), a suíte de ferramentas Express (gratuitas) da Microsoft foi empregada, portanto: o RDB utilizado foi o Microsoft SQL Server

2008 R2 Express; a linguagem de programação utilizada foi CSharp (c#); ambiente para desenvolvimento o Visual Studio 2010 Express [13].

O computador utilizado no estudo possuí processador Intel® Core™ 2 Duo CPU P7550, velocidade de clock 2.26GHz, memória RAM de 4GB e sistema operacional Microsoft Windows 7 Ultimate 64bits com Service Pack 1 instalado.

Resultados

Foram criadas 34 tabelas no RDB: Anonymize, AnonymizeItems, AnonymizeStudy, Concept, ConceptSource, ConceptTerm, Context, Contextualization, Document, DocumentType, DocumentWord, Idiom, KnownTerminology, LogSys, NormalizationType, NormalizationTypeStudy, OntologyConcept, OntologyPositionStudy, OntologyRelation, PeopleContext, PeopleStudy, StopWord, StopWordStudy, Study, StudyKnownTerminology, StudyProfile, StudyStatistics, StudyTerminology, Term, TermDocument, TermSource, TermTextualCategory, TextConvert e WordText.

Para facilitar o mapeamento objeto-relacional as classes da camada de acesso a dados recebem o mesmo nome das tabelas e as classes da camada de negócio são especializações das classes de acesso a dados e recebem a sufixo "BLL" (Business Logic Layer), como ilustrado na figura 2.

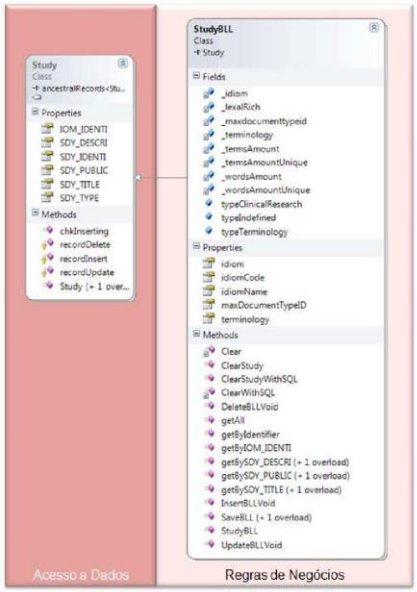


Figura 2: Ilustração da especialização da classe Study (estudo) entre as camadas de acesso a dados e a camada de regras de negócio (BLL).

Para a análise da geração dos termos candidatos com gramatura "1" (nGram=1) por normalização lexical foi criado o gráfico 1. O gráfico é montado em escala logarítmica para permitir à apresentação de 5 (cinco) variáveis com unidades diferentes simultaneamente e soma as variáveis das duas instituições privada e acadêmica. As variáveis são: "Qtd. Termos Candidatos Distintos" exibe a quantidade de termos candidatos únicos (que não se repetem) gerados após a normalização lexical; "Qtd. Unidades Lexicais Distintas" mostra a quantidade de unidades lexicais únicas (que não se repetem) que foram processadas; "Qtd. Unidades Lexicais" mostra a quantidade de unidades lexicais que foram processadas; "Qtd. Registros no banco de dados" é a quantidade de linhas das tabelas, ou seja, os registros gerados após a normalização; "Tempo em minutos" tempo gasto para concluir cada normalização lexical.

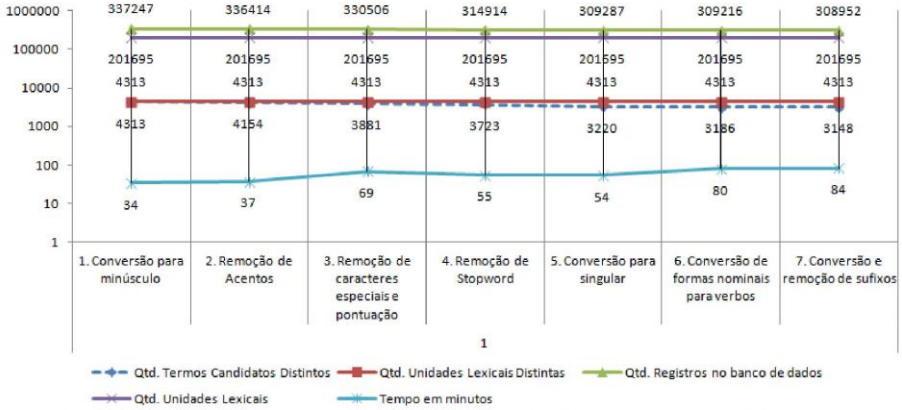
No gráfico 1, as variáveis "Qtd. Unidades Lexicais Distintas" e "Qtd. Unidades Lexicais" permanecem inalteradas após as normalizações lexicais. Isto acontece porque as unidades lexicais servem de base para geração dos termos candidatos, mas não são alteradas.

Na conversão para minúsculo, como era de se esperar, não houve variação entre a quantidade termos candidatos distintos e a quantidade de unidades lexicais distintas porque a

mudança para minúsculo não altera a forma de escrita do léxico. Para esta normalização lexical foram consumidos 34 minutos de processamento e geradas 337.247 registros no RDB.

Ao aplicar à remoção de acentos, 159 unidades lexicais se aglutinaram para formar 4.154 termos candidatos em 37 minutos de processamento armazenando 336.414 registros no RDB.

Gráfico 1: Análise da geração de termos candidatos em escala logarítmica (ngram=1)



Os picos de consumo de tempo ocorreram na remoção de caracteres especiais e pontuação, conversão para singular, conversão de formas nominais para verbos e conversão e remoção de sufixo e a explicação para isso são duas: 1) o algoritmo dessas normalizações precisou analisar o final das unidades lexicais ficando mais complexos que nas outras normalizações onde houve praticamente substituição textual. 2) o aumento gradual da quantidade de registros do RDB influenciou no tempo de processamento.

Foram conceituados automaticamente através do DeCS 207 (duzentos e sete) termos candidatos sendo 130 exatamente iguais aos termos candidatos procurados e 39 incompatíveis. Ontologia é a explicitação formal do conhecimento e, o laudo radiológico (prova textual) é um objeto do mundo real conhecido, portanto pode ser explicitado por ontologia. Ele apresenta uma estrutura esquemática global (superestrutura) com pelo menos duas categorias textuais (campos informacionais) bem definidas: "descrição" e "conclusão". Portanto, "descrição" e "conclusão" são categorias textuais que fazem parte da superestrutura textual chamada "laudo radiológico" e nelas se encontram os termos. No início da categoria textual "descrição", no primeiro parágrafo, se descreve o protocolo utilizado para coleta da imagem médica e, portanto, trata-se de uma categoria textual implícita chamada "protocolo". Para explicitar uma categoria textual implícita vinculada a uma posição vertical relativa dentro de outra categoria textual utiliza-se ontologia posicional vertical. Além de explicitar que o "protocolo" é uma categoria textual é necessário a relacionar como parte da categoria textual "descrição" porque ela existe no primeiro parágrafo desta categoria textual, ilustrado na figura 3. Isso permite aglutinar os

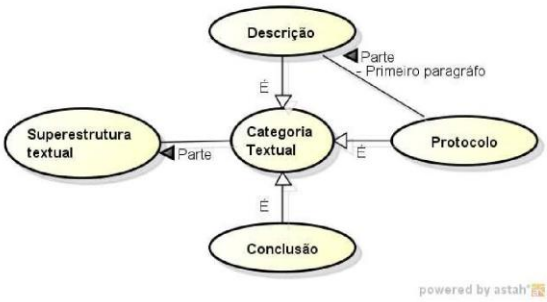


Figura 3: Ontologia posicional no laudo radiológico

termos dentro de cada categoria textual, de tal forma, que é possível listá-los. É possível identificar um conceito composto por outros conceitos utilizando ontologia posicional horizontal. Por exemplo, quando o conceito "grandeza" é precedido do conceito "número" (exemplo, "1 cm") é possível identificar o conceito de "medida".

Discussão

A utilização de camadas na construção da plataforma *generalis* resultou em regras de negócios testadas antes de serem utilizadas na camada de interface além de independência desta camada. Na prática, é possível construir outros tipos de interfaces (para celulares, por exemplo) utilizando o núcleo da plataforma *generalis* sem qualquer modificação porque o núcleo foi totalmente encapsulado em uma biblioteca de classe podendo ser consumido por outros projetos que utilizem o padrão Microsoft.Net.

Para criação de ferramentas computacionais que permitam construir produtos terminológicos dinâmicos e colaborativos faz-se necessário a utilização de ontologias e tecnologias que permitam relacionar os dados, como por exemplo, RDB. Existem esforços direcionados nesse sentido como o e-Termos [14], WordNet.br, VerbNet.br [15].

O Microsoft SQL Server se mostrou performático armazenando 2.246.536 registros em 413 minutos (aproximadamente 6 horas e 50 minutos), gerando um registro a cada 0,011 segundo. Como o mapeamento de toda a terminologia está na base de dados é possível recuperar informações utilizando a linguagem de consulta estruturada SQL (Structure Query Language) própria de RDB a qualquer momento sem a necessidade de reprocessamento. Por conta da velocidade e facilidade é mais que viável, é desejável, que as ferramentas de mineração de texto e construção de produtos terminológicos utilizem RDB ao invés de arquivos. Da primeira até última normalização lexical foram aglutinados 1.165 unidades lexicais em 3.148 termos candidatos totalizando 27% de aglutinação.

Ao utilizar ontologia posicional vertical foi possível listar os termos candidatos contidos na categoria textual "descrição", os termos candidatos da categoria textual "conclusão" e os termos candidatos da categoria textual "protocolo". Utilizando ontologia posicional horizontal é possível identificar um conceito formado pela composição de outros conceitos. No exemplo apresentado, ao identificar o conceito "medida", seria possível conhecer o aumento ou a diminuição dos nódulos de um paciente.

Conclusão

A plataforma *generalis* se apresentou capaz de extrair os termos candidatos e conceituá-los utilizando busca automática do conceito junto ao DeCS. As ontologias, inclusive as posicionais, foram criadas e armazenadas em banco de dados relacional.

Agradecimentos

Aos radiologistas e assistentes do Centro de Ciências das Imagens e Física Médica (CCIFM) do HCFMRP-USP, CEDIRP, Omnimagem e CIAw@re.

Em especial aos meus professores Dra. Maria Cristiane Barbosa Galvão e Dr. Valdair Francisco Muglia e ao doutor José Daniel Vieira de Castro.

Referências

[1] Krause D. Tópicos em Ontologia Analítica. Departamento de filosofia da Universidade Federal de Santa Catarina. 2011. Disponível em: http://www.cfh.ufsc.br/~dkrause/pg/TopicosOntologia.pdf. Acesso em 14 jun. 2012.

[2] Cimiano P, Reyle U, Saric J. Ontology-driven discourse analysis for information extraction. Data & Knowledge Engineering. 2005;55(1):59-83.

[3] Mhiri S, Despres S. Ontology usability via a visualization tool for the semantic indexing of medical reports (DICOM SR). In Proceedings of the 3rd Human-computer interaction and usability engineering of the Austrian computer society conference on HCI and usability for medicine and health care. 2007.

[4] Bellatreche L, Ait-Ameur Y, Chakroun C. A design methodology of ontology based database applications. Logic Journal of IGPL Advance Access. 2010.

[5] Gruber TR. Towards principles for the design of ontologies used for knowledge sharing. In: Guarino N, Poli R (eds.) Formal ontology in conceptual analysis and knownledge representation, Kluwer Academic Publishers, Dordrecht. 1993.

[6] Perna CL, Delgado HK, Finatto MJ. Linguagens especializadas em corpora: modos de dizer interface de pesquisa. EDIPUCRS, 2010. Disponível em http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/13706/8164. Acesso em 18 de jun. 2012.

[7] Kong Z, Zhang K. Summarizing Keyword Search in Relational Database. Third International Conference on Intelligent Human-Machine Systems and Cybernetics. 2011:152-155.

[8] BIREME, OPAS, OMS. DeCS: Descritores em ciências da saúde. [citado 2012 jun 21]. Disponível em http://decs.bvs.br.

[9] DeCS. Metodologias e aplicações. Disponível em http://decs.bvs.br/I/DocumentosDeCS_i.htm). Acesso em 20 de jun. 2012.

[10] Chiba C, Nardi A. Desenvolvimento em camadas. MSDN Brasil, 2007. Disponível em: http://www.microsoft.com/brasil/msdn/tecnologias/arquitetura/Layers_Developing.mspx. Acesso em 20 de jun. 2012.

[11] Microsoft Developer Network (MSDN). Anatomy of a Unit Test. Disponível em: http://msdn.microsoft.com/en-us/library/ms182517. Acesso em 13 de mai. 2012.

[12] DBDesigner. DBDesigner Fork 2009. Disponível em http://sourceforge.net/projects/dbdesigner-fork. Acesso em 18 de jun. 2012.

[13] Microsoft Visual Studio Express. Disponível em http://www.microsoft.com/visualstudio/express. Acesso em 16 de jun. 2012.

[14] Oliveira LHM. e-Termos: um ambiente colaborativo web de gestão terminológica. Tese (Doutorado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. 2009.

[15] Scarton CE. VerbNet.Br : construção semiautomática de um léxico computacional de verbos para o português do Brasil. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. 2011.

Contato

Flavio Barbosa,
http://www.flaviobarbosa.com.br
E-mail e Gtalk: flavio2020@gmail.com

9.3 ANEXO C: "Generalis: uma plataforma para explicitar o conhecimento" (Artigo publicado e apresentado oralmente durante o SIC2012)

Generalis: uma plataforma para explicitar o conhecimento

Flavio Barbosa
Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP)
flavio2020@gmail.com

Valdair Francisco Muglia
Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP)
fmuglia@fmrp.usp.br

Resumo: **O nome "*generalis*" é uma referência a um conceito da filosofia denominado metaphysica *generalis* que se ocupa da natureza e constituição da realidade e que foi chamada mais tarde de ontologia. No contexto de tecnologia da informação, vários artigos – Cimiano *et al*, 2005; Mhiri e Despres, 2007; Bellatreche *et al*, 2010 - definem ontologia como uma especificação explícita de uma conceitualização. "Generalis" é uma plataforma em ambiente WEB que permite a colaboração entre usuários de um determinado produto terminológico (estudo) e alimenta uma base única de conhecimento baseado em ontologia posicional. Ela é capaz de decompor documentos contendo texto não-estruturado em unidades lexicais utilizando técnicas de mineração de texto e os reconstruir em uma estrutura informacional conceitualizada utilizando ontológica posicional e apoiada por produtos terminológicos como o DeCS (descritores em ciências da saúde).** *Objetivo:* **Explicitar a estrutura informacional e o conhecimento de laudos radiológicos.** *Métodos:* **Coleta e análise de 1200 laudos radiológicos da coluna lombo sacra elaborados consecutivamente no primeiro trimestre de 2010 e primeiro trimestre de 2011 em duas instituições distintas sendo uma privada e outra pública acadêmica.** *Resultado:* **Foram extraídos 3142 termos candidatos distribuídos em 3 categorias textuais "descrição", "conclusão" e "protocolo", sendo esta última implícita, com percentual de densidade léxica de 72,5%. Automaticamente, do DeCS, foram recuperados 207 conceitos dos quais 130 exatamente iguais aos termos candidatos extraídos.** *Conclusão:* **A plataforma "generalis" se mostrou capaz de extrair os termos candidatos e os conceituar formalizando o conhecimento através da utilização de ontologia posicional.**

Palavras-chave: ***Ontologia posicional, Terminologia, Mineração de Texto, Laudo Radiológico e Texto Livre.***

INTRODUÇÃO

O nome "*generalis*" é uma referência a um conceito da filosofia denominado *metaphysica generalis* que se ocupa da natureza e constituição da realidade e que foi chamada mais tarde de ontologia (Krause, 2011). No contexto de tecnologia da informação, vários artigos – Cimiano *et al*, 2005; Mhiri e Despres, 2007; Bellatreche *et al*, 2010 - definem ontologia como uma especificação explícita de uma conceitualização (Gruber, 1993). Todavia, para entender o significado de conceitualização é necessário compreender as importantes diferenças entre dados, informação, conhecimento e inteligência.

Dados são informações sem significado para um ser humano. Imagine um número qualquer, por exemplo, o número 28 (vinte e oito) é um dado porque não representa coisa alguma neste momento.

Informação é um dado com significado para um ser humano, por exemplo: o ciclo menstrual de um determinado paciente é de 28 dias; como se sabe o significado do número 28 ele passa a ser uma informação.

Conhecimento é o ato ou efeito de conhecer. Conhecer é ter ou chegar a ter conhecimento, ideia, noção ou informação de alguma coisa ou de alguém. Ideia ou noção é sinônimo de conceito (Michaelis, 2012), portanto, conhecimento é o ato ou efeito de conceituar. Se ontologia é a especificação explícita de uma conceitualização, então é através dela que será possível formalizar o conhecimento, ou seja, ontologia tem a capacidade de formalizar e explicitar o conhecimento. No exemplo "o ciclo menstrual de um determinado paciente é de 28 dias" os dois conceitos principais são encontrados no DeCS (descritores em ciências da saúde) (BVS, 2012): "ciclo menstrual" é o período, em mulher ou fêmea primata com ovulação, que vai desde o início até a próxima hemorragia menstrual (...); "pacientes" são os indivíduos participantes do sistema de cuidados de saúde com o propósito de receber procedimentos terapêuticos, diagnósticos ou preventivos.

Inteligência é a faculdade de entender, pensar, raciocinar, interpretar; compreensão, conhecimento profundo (Michaelis, 2012). Ainda no exemplo: "o ciclo menstrual de um determinado paciente é de 28 dias", um pesquisador poderia solicitar a um sistema de informação hospitalar (HIS, do inglês *Hospital Information System*) que possua estes conceitos e suas relações, ou seja, possua uma ontologia formalizada e ativa, a busca do termo "ciclo menstrual" e, apenas os pacientes do sexo feminino seriam listados porque o SI interpretou a ontologia e identificou a relação existente entre os conceitos "ciclo menstrual" e "mulher ou fêmea" transmitindo a sensação de inteligência para o pesquisador.

Conceitualização, portanto, é o ato de atribuir significado aos dados através de relações apropriadas ("é", "parte de", "antagonista de", "etc.") com os respectivos conceitos do contexto de estudo a fim de transformar dados em conhecimento. Por este motivo, é necessário o estudo das palavras (léxicos) que compõe os documentos de uma determinada comunidade, por exemplo, as observações,

evoluções e laudos elaborados por uma determinada especialidade médica, com o objetivo de explicitar o conhecimento através do entendimento terminológico.

Existem duas definições para terminologia: 1) enquanto atividade é o conjunto vocabular, conjunto de léxicos, próprio de uma ciência; 2) enquanto teoria é o conjunto de postulados teóricos necessários para dar suporte à análise de fenômenos linguísticos concernentes à comunicação especializada. Para uma comunicação especializada é necessário à utilização de uma linguagem especializada. Para isso ocorrer, é necessário que o conteúdo seja descrito de maneira precisa, com referência a um esquema científico determinado ou uma estrutura normativa estabelecida. Desta forma, o conteúdo torna-se objeto de conhecimento especializado (Perna *et al*, 2010).

Laudos radiológicos são documentos produzidos por comunidade profissional especializada (médicos radiologistas) que descrevem de maneira precisa os achados observados nas imagens médicas de acordo com o conhecimento existente em referencial teórico cientifico. Portanto, são documentos que utilizam linguagem especializada e são constituídos por unidades lexicais que referenciam conceitos da atividade médica radiológica. Ora, termos são unidades lexicais que representam conceitos (Perna *et al*, 2010), portanto os laudos radiológicos são documentos constituídos de termos especializados da comunidade radiológica. Ontologia é a explicitação formal do conhecimento e, o laudo radiológico (prova textual) é um objeto do mundo real conhecido, portanto pode ser explicitado por ontologia.

As unidades lexicais são constituídas por uma ou mais palavras e, portanto, possuem uma espécie de gramatura. Quando uma unidade lexical é constituída por apenas uma palavra a gramatura é "1" e a notação é escrita da seguinte forma "nGram=1". Se uma unidade lexical é constituída de duas palavras (sintagma nominal) sua gramatura é "2" e a notação é escrita da seguinte forma "nGram=2" e recebe a denominação de bigrama. Quando a unidade lexical é constituída por três palavras a gramatura é "3" e a notação é "nGram=3" e recebe a denominação de trigrama e assim sucessivamente (Perna *et al*, 2010).

Uma mesma unidade lexical pode sofrer variações quanto à forma, por exemplo: singular e plural, dispersando a frequência daquela unidade lexical. Para evitar esse efeito, as unidades lexicais são transformadas para uma mesma forma lexical, ou seja, elas são normalizadas. Quando uma unidade lexical é normalizada ela se torna um termo candidato porque será com base nessa unidade lexical normalizada que se analisará uma associação com um conceito. Quando o conceito for associado ao termo candidato lhe é atribuído significado e, portanto, passará a ser um termo efetivamente. Existem unidades lexicais que possuem pouca ou nenhuma chance de se transformarem em termos porque não representam um conceito. Elas existem por conta da sintaxe da linguagem ou para atribuir sentido a frase, à lista dessas unidades lexicais dá-se o nome de "stopwords". As "stopwords" não geram termos candidatos e, portanto, nunca se transformarão em termos, serão desconsideradas. São exemplos de "stopwords" utilizadas nesse estudo: "a", "acaso", "ai" e "foi" (Rezende *et al*, 2011).

METODOLOGIA

Os documentos utilizados são laudos radiológicos da coluna lombo sacra elaborados consecutivamente no primeiro trimestre de 2010 e primeiro trimestre de 2011. Para formar o corpus textual a ser analisado, foram coletados 600 laudos radiológicos de uma instituição privada e outros 600 de uma instituição pública acadêmica totalizando 1200 laudos. Nestes laudos não havia qualquer informação que pudesse identificar o paciente, médico executante ou médico solicitante e, portanto, são anônimos. As análises apresentadas neste trabalho serão quantitativas. O trabalho foi submetido ao comitê de ética em pesquisa clínica e recebeu o número CAAE 00593812.2.1001.5440.

As seguintes normalizações lexicais foram aplicadas com a finalidade de gerar os termos candidatos: 1) conversão para minúsculo; 2) remoção de acentos; 3) remoção de caracteres especiais e de pontuação; 4) remoção de stopword; 5) conversão de plural para singular; 6) conversão de formas nominais para verbos; 7) conversão e remoção de sufixos. Embora a plataforma *generalis* permita a geração de termos candidatos com gramatura igual a cinco (nGram=5), neste estudo as observações são referentes a gramatura "1" (nGram=1). Para cada tipo de normalização lexical foi criado um estudo específico dentro da plataforma *generalis* com o objetivo de observar o efeito de cada normalização lexical. Para avaliar o efeito se empregou o indicador "densidade léxica" (DL) porque ele avalia o texto em termos da variedade e quantidade de vocabulário existente calculando a proporção entre o número de termos candidatos (unidades lexicais normalizadas = Nlex) e o total de unidades lexicais (palavras = N) do conjunto de documentos (corpus textual) de cada instituição (Roberto *et al*, 2012). Quanto maior a DL, maior é a variedade de léxicos de um documento ou de um conjunto de documentos.

$$\text{Densidade Léxica (DL)} = \frac{Nlex}{N}$$

A arquitetura de software empregada para plataforma *generalis* é de camadas (Chiba e Nardi, 2007), como segue: acesso a dados, regra de negócios, testes unitários (MSDN, 2012) e interface. A figura 1 ilustra a relação entre o banco de dados relacional (RDB – Relational database) (Kong e Zhang, 2011) e as camadas da plataforma.

Figura 1: Arquitetura de software da plataforma *generalis*.

Com a exceção da ferramenta DBDesigner (DBDesigner, 2009) para desenho do modelo de entidade e relacionamento (MER – Model Entity Relation), a suíte de ferramentas Express (gratuitas) da Microsoft foi utilizada, portanto: o RDB utilizado foi o Microsoft SQL Server 2008 R2 Express; a linguagem de programação utilizada foi CSharp (c#); ambiente para desenvolvimento o Visual Studio 2010 Express (MS-VSE, 2012).

No momento da análise dos termos candidatos para conceitualização foi solicitada à plataforma *generalis* a geração automática dos conceitos utilizando um produto terminológico conhecido: o DeCS (descritores em ciências da saúde) (BVS, 2012). Para este estudo duas consultas foram feitas ao DeCS usando o idioma português: uma por unidade lexical exata e outra por unidade lexical semelhante.

RESULTADOS E DISCUSSÃO

As tabelas do modelo de entidade e relacionamento (MER) desenhadas no DBDesigner foram aplicadas no RDB criando 34 tabelas.

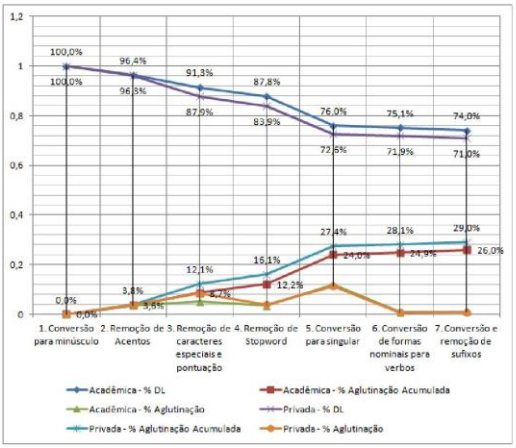


Gráfico 1: Percentual de Densidade Léxica (DL*100) por normalização lexical

Foram extraídos 1949 termos candidatos na instituição acadêmica e 1193 da instituição privada. O gráfico 1 está em escala percentual para permitir comparar o efeito das normalizações nas duas instituições simultaneamente porque as quantidades de termos candidatos entre as instituições privada e acadêmica são diferentes. Ele possibilita a verificação do efeito da aplicação de cada normalização lexical através da observação do percentual de DL.

Na conversão para minúsculo, como era de se esperar, não houve variação entre a quantidade de termos candidatos e a quantidade de unidades lexicais. A DL se manteve em 100% (cem por cento) porque a mudança para minúsculo não altera léxico.

Ao aplicar à remoção de acentos ocorre a primeira aglutinação (condensação) das unidades lexicais em termos candidatos em 3,6% na instituição acadêmica e 3,7% na instituição privada. Aplicando a remoção de caracteres especiais e pontuação houve uma aglutinação de 5,0% na instituição acadêmica e 8,4% na instituição privada. Aplicando a remoção de "stopwords" ocorreu aglutinação de 3,5% na instituição acadêmica e 3,9% na instituição privada. Na conversão do plural para singular houve 11,85% de aglutinação na instituição acadêmica e 11,37% na instituição privada. A conversão de formas nominais em verbos resultou em 0,9% de aglutinação na instituição acadêmica e 0,6% de aglutinação na instituição privada. A remoção de sufixo resultou em 1,1% de aglutinação na instituição acadêmica e 0,5% na instituição privada.

Após a última normalização, remoção de sufixo, o percentual de DL para instituição acadêmica ficou em 74,0% e com percentual de aglutinação acumulado de 26% e, para instituição privada, o percentual de DL ficou em 71,0% e com aglutinação acumulada de 29%. Esse resultado indica que há menor variação de termos candidatos na instituição privada do que na instituição acadêmica. O percentual médio de DL das instituições ficou em 72,5% com aglutinação acumulada média de 27,5%.

Foram conceituados automaticamente através do DeCS 207 (duzentos e sete) termos candidatos sendo 130 exatamente iguais aos termos candidatos procurados e 39 incompatíveis (vide exemplo na tabela 1).

Compativel	Termo Candidato	DeCS	
		Termo	Definição
Não	cabeca	Face	Parte anterior da cabeça que inclui a pele, os músculos e as estruturas da fronte, olhos, nariz, boca, bochechas e mandíbula.
Sim	gota	Gota	Doença metabólica hereditária caracterizada por artrite aguda recorrente, hiperuricemia e deposição de uratos de sódio dentro e ao redor das articulações com a formação de cálculos de ácido úrico.
Sim	infarto	Infarto	Formação do infarto, que é NECROSE no tecido, causado por ISQUEMIA local devido a obstrução da CIRCULAÇÃO SANGUÍNEA (geralmente por trombo ou êmbolo).

Tabela 1: Exemplo de termos recuperados do DeCS.

Identificou-se no laudo radiológico uma estrutura esquemática global (superestrutura) com pelo menos duas categorias textuais (campos informacionais) bem definidas: "descrição" e "conclusão". Portanto, "descrição" e "conclusão" são categorias textuais que fazem parte da superestrutura textual chamada "laudo radiológico" onde se encontram os termos.

		Tipo da Instituição			
		Acadêmica		Privada	
Total de Termos	Qtd. Termos Candidatos	1949	100%	1193	100%
Categorias Textuais	Termos distintos na Descrição	1716	88%	1085	91%
	Exclusivamente na Descrição	1048	54%	757	63%
	Termos distintos na Conclusão	899	46%	433	36%
	Exclusivamente na Conclusão	231	12%	105	9%
	Termos distintos no Protocolo			18	1%
	Exclusivamente no Protocolo			0	0%

Tabela 2: Distribuição dos termos entre as categorias textuais

Na instituição privada, o primeiro parágrafo da categoria textual "descrição" trata do protocolo utilizado para coleta da imagem médica e, portanto, trata-se de uma categoria textual implícita chamada "protocolo". Para explicitar uma categoria textual em uma posição vertical relativa ou fixa dentro de outra categoria textual, utiliza-se ontologia posicional vertical. Além de explicitar que o "protocolo" é uma categoria textual é necessário a relacionar com o primeiro parágrafo da categoria textual "descrição", ilustrado na figura 2, permitindo o agrupamento de termos por categoria textual.

Analisando a distribuição dos termos entre as categorias textuais na tabela 2, verifica-se que em média apenas 10,5% dos termos são utilizados exclusivamente na categoria textual "conclusão" e que em média 58,5% são utilizados exclusivamente na categoria textual "descrição". A categoria textual "protocolo" foi identificada apenas na instituição privada e, por fazer parte da categoria textual "Descrição" não possuí termos exclusivos.

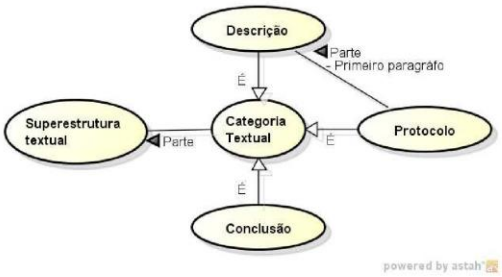


Figura 2: Ontologia posicional vertical da superestrutura "laudo radiológico"

Um conceito pode ser composto ou determinado por outros conceitos. Por exemplo, quando o conceito "unidade de medida" é precedido do conceito "número" (exemplo, "1 cm") o conceito "medida" é identificado. Para conceituar "medida" é necessário utilizar ontologia posicional horizontal porque ela permite identificar um conceito analisando a posição horizontal dos termos.

CONCLUSÃO

Extrair termos candidatos não é extrair conhecimento porque é preciso relacioná-los apropriadamente com os seus respectivos conceitos. A plataforma *generalis* se mostrou capaz de extrair e conceituar os termos candidatos os relacionando apropriadamente com os seus respectivos conceitos e estes entre si, utilizando ontologia posicional.

REFERÊNCIAS

KRAUSE, D. Tópicos em Ontologia Analítica. Departamento de filosofia da Universidade Federal de Santa Catarina, jun. 2011. Disponível em http://www.cfh.ufsc.br/~dkrause/pg/TopicosOntologia.pdf. Acesso em: 14 jun. 2012.

CIMIANO, P.; REYLE, U.; SARIC, J. Ontology-driven discourse analysis for information extraction. Data & Knowledge Engineering v. 55, 2005.

MHIRI, S.; DESPRES, S. Ontology usability via a visualization tool for the semantic indexing of medical reports (DICOM SR). In Proceedings of the 3rd Human-computer interaction and usability engineering of the Austrian computer society conference on HCI and usability for medicine and health care, nov. 2007.

BELLATRECHE, L.; AIT-AMEUR, Y.; CHAKROUN C. A design methodology of ontology based database applications. Logic Journal of IGPL Advance Access, 2010.

GRUBER, T.R. Towards principles for the design of ontologies used for knowledge sharing. In: Guarino, N., Poli, R. (eds.) Formal ontology in conceptual analysis and knownledge representation, Kluwer Academic Publishers, Dordrecht, 1993.

MICHAELIS. Moderno Dicionário da Língua Portuguesa. Disponível em http://michaelis.uol.com.br/moderno/portugues/index.php. Acesso em: 14 jun. 2012.

BIBLIOTECA Virtual em Saúde. Descritores em Ciências da Saúde. São Paulo: BVS, 2012. Disponível em http://decs.bvs.br. Acesso em: 29 mar. 2012.

PERNA, C.L.; DELGADO, H.K.; FINATTO, M.J. Linguagens especializadas em corpora: modos de dizer interface de pesquisa. EDIPUCRS, 2010. Disponível em http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/13706/8164. Acesso em: 18 jun. 2012.

REZENDE, S.O.; MARCACINI, R.M.; MOURA M.F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. *Revista de Sistemas de Informação da FSMA,* 2011.

ROBERTO, J.A.; ANTÒNIA, M.M.; SALAMÓ M. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. Revista Procesamiento de Lenguaje Natural, n. 48, 2012.

CHIBA, C.; NARDI, A. Desenvolvimento em camadas. MSDN Brasil, 2007. Disponível em http://www.microsoft.com/brasil/msdn/tecnologias/arquitetura/Layers_Developing.mspx. Acesso em: 20 jun. 2012.

MSDN. Microsoft Developer Network (MSDN). Anatomy of a Unit Test. Disponível em http://msdn.microsoft.com/en-us/library/ms182517. Acesso em: 13 mai. 2012.

KONG, Z.; ZHANG, K. Summarizing Keyword Search in Relational Database. Third International Conference on Intelligent Human-Machine Systems and Cybernetics, 2011.

DBDESIGNER. DBDesigner Fork 2009. Disponível em http://sourceforge.net/projects/dbdesigner-fork. Acesso em: 18 jun. 2012.

MS-VSE. Microsoft Visual Studio Express. Disponível em http://www.microsoft.com/visualstudio/express. Acesso em: 16 jun. 2012.

AGRADECIMENTOS

Aos radiologistas e assistentes do Centro de Ciências das Imagens e Física Médica (CCIFM) do HCFMRP-USP, CEDIRP, Omnimagem e CIAw@re.

Em especial aos meus professores Dra. Maria Cristiane Barbosa Galvão e Dr. Valdair Francisco Muglia e ao doutor José Daniel Vieira de Castro.

COMO CITAR ESTE ARTIGO

BARBOSA, Flávio; MUGLIA, Valdair Francisco. *Generalis*: uma plataforma para explicitar o conhecimento. In: SIMPÓSIO SOBRE INFORMAÇÃO CLÍNICA, Ribeirão Preto, 2012. Anais... Ribeirão Preto : HCFMRP, 2012. Disponível em: http://www.hcrp.usp.br/sic2012.

9.4 ANEXO D: "Metodologia para estruturação de informações de laudos radiológicos em português." (Artigo a ser publicado)

Metodologia para estruturação de informações de laudos radiológicos em português.

Flavio Barbosa

Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP)
contato@flaviobarbosa.com.br

Maria Cristiane Barbosa Galvão

Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP)
mgalvao@fmrp.usp.br

Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP)
fmuglia@fmrp.usp.br

Valdair Francisco Muglia

RESUMO

Introdução: O laudo radiológico com entrada de dados estruturada proporciona precisão na recuperação de informações, todavia, é menos abrangente que o texto-livre e pode limitar o especialista de descrever um achado importante não contemplado pela estrutura pré-definida. Portanto, a solução está em interpretar o texto escrito pelo especialista e os transportar para uma estrutura computacional ontológica.

Objetivos: Elaborar uma metodologia para estruturação de informações de laudos radiológicos que contemple as especificidades brasileiras, incluindo a terminologia a ser empregada.

Métodos: Coleta de 1701 laudos radiológicos de ressonância magnética da coluna lombossacra em três instituições distintas. Para estruturar a informação foram utilizadas técnicas de mineração de texto e conceituação ontológica das unidades lexicais. Dez radiologistas especialistas avaliaram a superestrutura textual e a terminologia extraída através de questionário eletrônico.

Resultados: Foram identificadas três superestruturas textuais diferentes com variação terminológica na denominação de suas categorias textuais. A quantidade de termos candidatos conceituada foi de 4183, originando 727 conceitos em português, 13963 relações entre termos candidatos e conceitos e 789 relações entre conceitos. A metodologia criada é constituída por seis passos: 1) coleta dos laudos radiológicos de determinado tipo de exame; 2) decomposição textual; 3) normalização das unidades lexicais; 4) identificação das superestruturas textuais; 5) conceituação dos termos candidatos; 6) avaliação das superestruturas identificadas e da terminologia extraída pelos especialistas na linguagem de especialidade estudada.

Conclusão: A metodologia utilizada permitiu a estruturação da informação porque foi capaz de indicar as superestruturas textuais, extrair as unidades lexicais, as normalizar e conceituar ontologicamente mantendo as referências com as suas respectivas categorias textuais e laudos radiológicos.

Descritores: Laudo radiológico, texto livre, estruturação da informação, ontologia, terminologia.

ABSTRACT

Introduction: The radiological report with structured data entry provides precision in information retrieval, however, is less ample than the text-free and may limit the specialist to describe an important finding not covered by pre-defined structure. So the solution is to interpret the text written by specialist and carry for a computational ontological structure.

Objectives: Develop a methodology for structuring information in radiology reports covering the Brazilian specifications, including the terminology to be used.

Methods: Was collected 1701 radiological reports of MRI of the lumbosacral spine in three different institutions. To structure the information were used techniques of text mining and ontological conceptualization of lexical units extracted. Ten radiologists specialists evaluated the textual superstructure and terminology extracted by the electronic questionnaire.

Results: Identified three different textual superstructures with terminological variation in the name of their textual categories. The amount of candidate terms was conceptualized 4183, yielding 727 concepts in Portuguese, 13963 relations between candidate terms and concepts and 789 relations between concepts. The established methodology consists of six steps: 1) collection of radiology reports of certain type of examination; 2) decomposition textual; 3) normalization of lexical units; 4) identification of textual superstructures; 5) conceptualization of candidate terms; 6) evaluation superstructures identified and terminology extracted by experts in the specialized language study.

Conclusions: The methodology allowed the structuring of information because it was able to indicate the textual superstructures, extracting lexical units, the normalizing and conceptualize ontologically while maintaining references to their respective categories and textual radiology reports.

Uniterms: Radiological report, free-text, structuring information, ontology, terminology.

INTRODUÇÃO:

Uma língua é um sistema de sinais apropriados a uma notação (MICHAELIS, 2009). Os sinais, no universo cultural de uma comunidade linguística, equivalem ao conjunto de palavras disponíveis no idioma e formam o seu léxico. O léxico de uma língua é composto por um sistema de unidades, dito unidades lexicais, que carregam consigo o conhecimento geral desta comunidade (Reis, 2010). Portanto, a linguagem comum ou língua geral é o sistema de comunicação oral e escrita de uso cotidiano em contexto geral de uma comunidade linguística (Pavel; Nolet, 2002).

Nas atividades humanas especializadas, que exigem longo período de aprendizado para seu conhecimento e experiência para a sua prática as unidades lexicais da língua geral também são especializadas pelos indivíduos que as executam atribuindo-lhes novo significado e, criando assim, os termos de uma língua especializada. A especialização da língua ocorre porque para exercer uma atividade específica a comunicação necessita ser diferenciada, ou seja, o seu conteúdo deve ser descrito de maneira precisa, sem ambiguidades, com referência a um esquema científico determinado ou uma estrutura normativa estabelecida (Maciel, 2010). Ao equacionar, mapear, identificar um determinado conhecimento o especialista o organiza mentalmente em um conceito. Portanto, o conceito é a organização mental de um conhecimento, uma unidade de pensamento ou conhecimento, e é concebido através de uma operação mental podendo ser expresso e, portanto compartilhado, na comunidade especializada através de uma unidade lexical especializada denominada "termo" (Silva-Filho, 2010). Todavia, o termo não é o conceito propriamente dito, mas sim, a chave de acesso, o referencial, que leva a mente humana a recuperar a organização mental de um conhecimento (Francelin; Kobashi, 2011).

A linguagem especializada é amplamente utilizada na área da saúde porque possui termos próprios, que geralmente referenciam um conceito único dentro da comunidade especializada em saúde que é usado, a fim de descrever precisamente as observações médicas referente ao que se observa nos exames laboratoriais, nas imagens médicas ou no exame clínico de determinado paciente.

As observações médicas são escritas em documentos que, não raro, possuem uma estrutura esquemática global que o caracteriza, diz-se, que ele possui uma superestrutura textual. A superestrutura textual é composta por áreas de texto bem definidas chamadas categorias textuais (Vieira; Correa, 2011). Um texto contendo as categorias textuais introdução, objetivo, metodologia, resultado, discussão, conclusão e referências bibliográficas; um pesquisador o reconhece como um artigo científico porque a superestrutura textual o identifica como tal. Da mesma forma, quando um médico recebe um documento contendo os dados demográficos do paciente, médico solicitante, técnica do exame, descrição e impressão diagnóstica e o médico executante; ele o reconhece como um laudo de exame (figura 1).

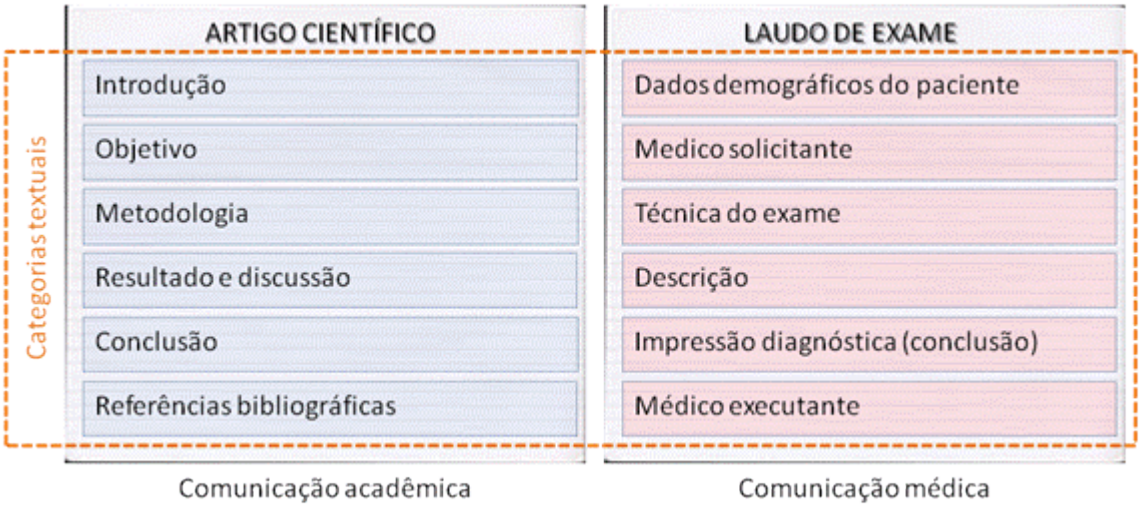


Figura 21: Exemplo de superestruturas textuais em diferentes atividades especializadas

No caso de um exame laboratorial, por exemplo, o médico que solicita é o mesmo que analisa e interpreta o resultado do exame, diferente do exame de diagnóstico por imagem que exige um médico executante especializado em análise e interpretação das imagens médicas coletadas para identificar o que elas revelam (Maciel-Lima; Rasia, 2012). O médico executante, chamado de radiologista ou imaginologista, é responsável por coordenar a execução do exame de diagnóstico por imagem e de registrar sua análise e interpretação em um relatório médico denominado "laudo radiológico". Geralmente, o laudo radiológico contém as seguintes categorias textuais: a técnica utilizada no exame, os dados do paciente, a região examinada, a descrição dos achados da imagem e, quando o diagnóstico é assertivo, uma conclusão

ou impressão diagnóstica (Fenelon, 2003; Prestes-Júnior; Tourinho; Rangel, 2012). Este relatório é enviado ao médico solicitante que concluirá ou continuará a investigação sobre o problema do paciente.

O laudo radiológico é o meio de comunicação pelo o qual o médico solicitante toma ciência das interpretações do médico executante onde a comunicação deve ser exata, livre de ambiguidades. A comunidade radiológica persegue a harmonização terminológica, um consenso para empregar o termo em determinada situação (Pavel; Nolet, 2002), e criou diversos produtos terminológicos ao longo do tempo como o índice ACR (American College of Radiology´s Index for Radiological Diagnoses), o RadLex da RSNA (Radiological Society of North America) e o BI-RADS (Breast Imaging Reporting and Data System) (ACR, 2003) que além da terminologia propõe uma estrutura para o exame de ultrassonografia de mama bem definida e mundialmente aceita inclusive no Brasil (Passman *et al.*, 2011).

A estruturação dos laudos radiológicos, em geral, se realiza de forma empírica perguntando a cada uma das equipes, solicitantes e executantes, quais as categorias textuais são necessárias para compor o laudo radiológico. Desta forma, informações ou termos que são frequentemente empregados podem deixar de serem considerados, provocando rejeição da utilização do laudo estruturado. No caso do Brasil, que possui dimensão territorial continental, médicos de regiões diferentes contestam termos de domínio pré-definidos (descritores) por estarem habituados com terminologia própria de sua região. A entrada de dados estruturada proporciona precisão na recuperação de informações, todavia, é menos abrangente que o texto-livre e pode limitar o médico executante de descrever um achado importante não contemplado pela estrutura pré-definida. Portanto, a solução está em interpretar o texto escrito pelo especialista e os transportar para uma estrutura computacional ontológica.

A proposta desta pesquisa é propor uma metodologia para estruturação de informações de laudos radiológicos que contemple as especificidades brasileiras, incluindo a terminologia a ser empregada.

MATERIAL E MÉTODOS

A amostra foi composta de 1701 (um mil e setecentos e um) laudos consecutivos de ressonância magnética de coluna lombossacra para três instituições: A) Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (HCFMRP-USP), acadêmica e no estado de São Paulo (SP); B) Clínica CEDIRP de Ribeirão Preto, privada e no estado de São Paulo (SP); C) Omnimagem de Fortaleza, privada e no estado do Ceará (CE).

Instituição	Tipo	Trimestre	Estado da União	Quantidade de Laudos
A	Acadêmica	jan/10	São Paulo	600
B	Privada	jan/10	São Paulo	600
C	Privada	jan/11	Ceara	501
	Total de laudos coletados			**1701**

Tabela 19: Descrição das amostras coletadas por instituição

O estudo foi submetido à apreciação do Comitê de Ética em Pesquisa do HCFMRP-USP, através da Plataforma Brasil do Ministério da Saúde e, foi aprovado por este comitê em 16/07/2012 sob o número CAAE 00593812.2.1001.5440.

Para acomodar o corpus textual e executar a mineração de texto e organização ontológica foi criada uma ferramenta computacional chamada *meta-genaralis* construída utilizando a linguagem c# e arquitetada em camada de acesso a dados, regra de negócios e testes unitários (figura 3).

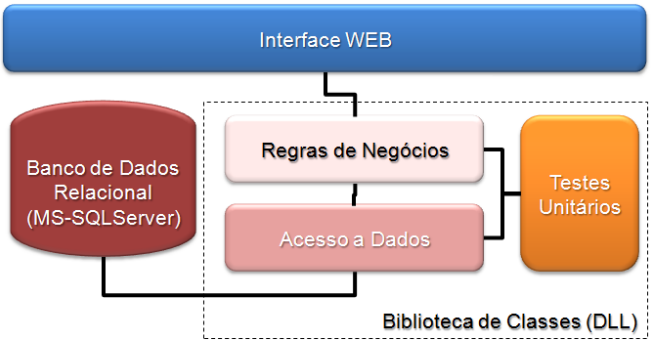


Figura 22: Arquitetura da ferramenta computacional *Meta-generalis*

Utilizando técnicas de mineração de texto, o corpus textual foi decomposto em parágrafos, frases e unidades lexicais ("tokens"). As unidades lexicais sofreram 7 normalizações: 1) conversão para minúsculo; 2) remoção de acentos; 3) remoção de caracteres especiais e de pontuação; 4) remoção de *stopwords*; 5) conversão de plural para singular; 6) conversão de formas nominais para verbos; 7) conversão de sufixos.

Para avaliar o impacto de cada normalização foram utilizados dois indicadores: a densidade léxica (DL) porque é definida como a razão entre a quantidade de unidades lexicais normalizadas (*Nlex*) e a quantidade de unidades lexicais total do corpus textual (*N*) (Roberto; Martí; Llorente, 2012) e a aglutinação léxica (AL) que é definida pelo valor complementar à densidade léxica.

$$DL = \frac{NLex}{N}$$

Equação 3: Índice de densidade léxica (DL)

$$AL = (1-DL)$$

Equação 4: Definição do índice de aglutinação léxica (AL)

O arquivo invertido, estrutura de dados que armazena as unidades lexicais aos seus respectivos documentos, foi armazenado no banco de dados relacional Microsoft SQL Server 2008 R2 Express (gratuito). Termo candidato composto, àqueles constituídos por mais de uma unidade lexical (nGram>1), foram identificados por decisão ontológica contextual horizontal (vide figura 3) que consiste em identificar o termo candidato levando em consideração o conceito anterior ou posterior. Por isto, os resultados deste estudo são referentes a termo simples (nGram=1).

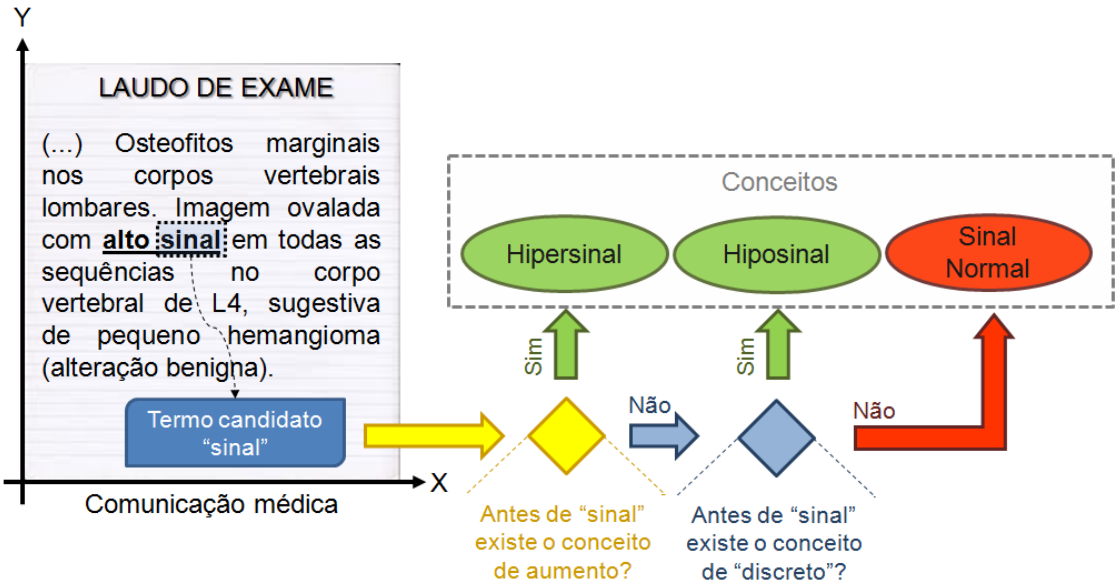


Figura 23: Exemplo de utilização da decisão ontológica contextual horizontal

A lista de termos candidatos referente ao corpus textual em estudo foi totalmente conceituada em uma estrutura ontológica seja de forma automática, utilizando produtos terminológicos eletrônicos (PTE), seja manualmente, através da literatura ou consultando especialistas da área. Foram empregadas as relações "É", "ASSOCIADO", "INDICA" e "PARTE DE".

Para identificar a superestrutura textual os termos candidatos que possuíam a frequência entre 95% até 100% da quantidade de documentos do estudo foram listados, pois os termos com frequência igual ou próxima a quantidade de documentos do corpus textual são termos candidatos a representarem categorias textuais. Os termos candidatos listados nesta condição foram analisados dentro do contexto que os originou e quando se tratava da identificação de uma categoria textual eles foram relacionados como objetos do conceito "categoria textual" dentro da ontologia.

Os termos candidatos foram recuperados do PTE de descritores de ciências da saúde em português (BIREME, 2007), por requisição via protocolo de transferência de hipertexto (HTTP) e interpretação da chave de definição ("<definition>") dentro do arquivo de linguagem de marcação (XML) retornado. Após a busca por definições dos termos candidatos no DeCS os conceitos retornados foram avaliados e identificados como compatíveis ou incompatíveis. Os termos candidatos que não possuíam referencia no PTE foram conceituados manualmente através da literatura ou consultando especialistas. As fontes literárias consultadas foram: dicionário on-line Michaelis (MICHAELIS, 2009) porque contém o conceito associado ao termo em diversas línguas de especialidade inclusive nas ciências da saúde, o Dorland dicionário médico ilustrado (Dorland, 1999), o sistema de linguagem médica unificada (UMLS, 2013), artigos científicos, órgãos de pesos e medidas internacionais e atlas de anatomia humana (Netter, 2000).

A conceitualização de todos os termos candidatos de um determinado estudo, dentro da ferramenta computacional *Meta-generalis*, transforma o próprio estudo em um PTE capaz de retornar o conceito de um determinado termo candidato, desta maneira, somente o estudo da primeira instituição utilizou apenas o DecS, os demais utilizaram o PTE criado pelos estudos das instituições anteriores.

Dez radiologistas especialistas, no mínimo três anos de experiência neste tipo de exame, avaliaram a superestrutura textual e a terminologia extraída através de questionário eletrônico sendo distribuídos da seguinte forma: seis da instituição "A", dois da instituição "B" e dois da instituição "C".

A lista de termos a ser avaliada pelos radiologistas especialistas foi criada utilizando os seguintes critérios:

1. Uma lista de termos que possuíam diferença lexical entre as instituições foi gerada, denominada lista1 contendo 372 (trezentos e setenta e dois) termos;

2. Com base na lista1, o orientador deste trabalho, que é radiologista especialista, excluiu os termos referentes à: anatomia humana (artéria, fígado, etc.), comuns à linguagem natural (correlação, complexo, etc.), descritores de doenças ou patologias de menor importância (cisto, lipoma, etc.) e procedimentos (exemplo: ultrassonografia, tomografia, etc.). Foram excluídos 95 termos e, portanto, a lista2 foi composta por 277 termos.

3. Usando a função "ALEATÓRIOENTRE" do Microsoft Excel um número aleatório foi atribuído a cada um dos 277 termos da lista2 e ordenados do maior para o menor.

4. Após esse processo, foram selecionados os 30 (trinta) primeiros termos para o questionário de avaliação.

RESULTADOS

Quantitativos:

A quantidade de unidades lexicais brutas extraídas das instituições A, B e C foi respectivamente de 2633, 1680 e 1816 totalizando 6129. Após o sétimo e, último, tipo de normalização a quantidade de unidades lexicais normalizadas nas instituições A, B e C foi respectivamente de 1949, 1193 e 1426 totalizando 4568 (vide figura 4).

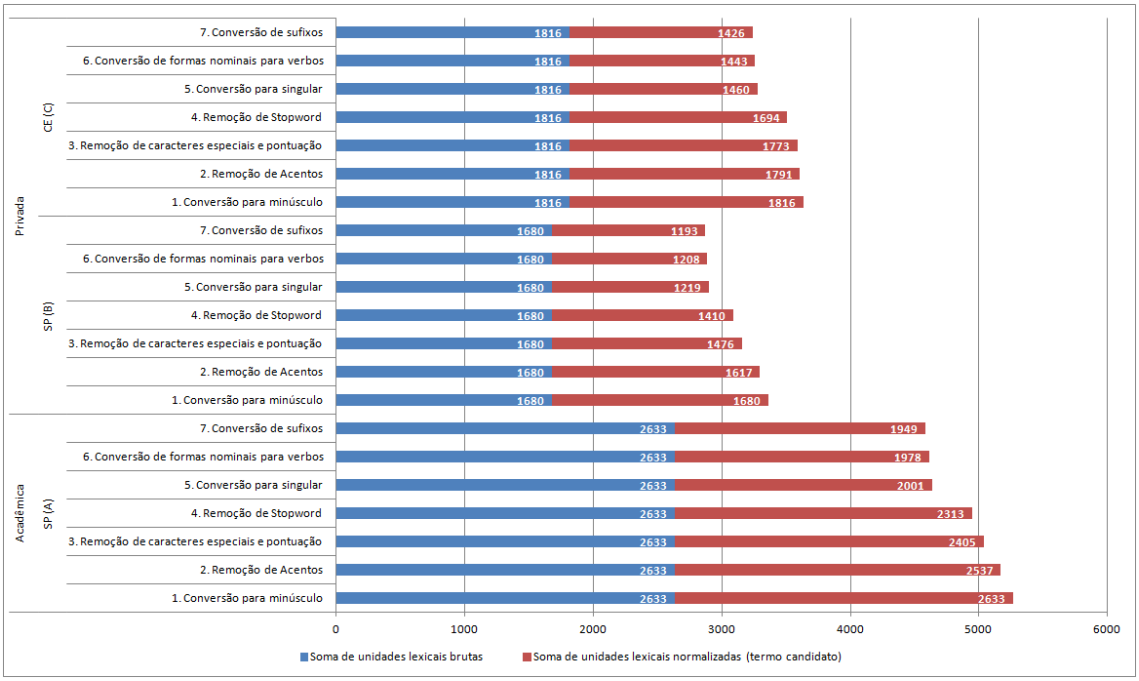


Gráfico 8: Resultado absoluto das normalizações por instituição

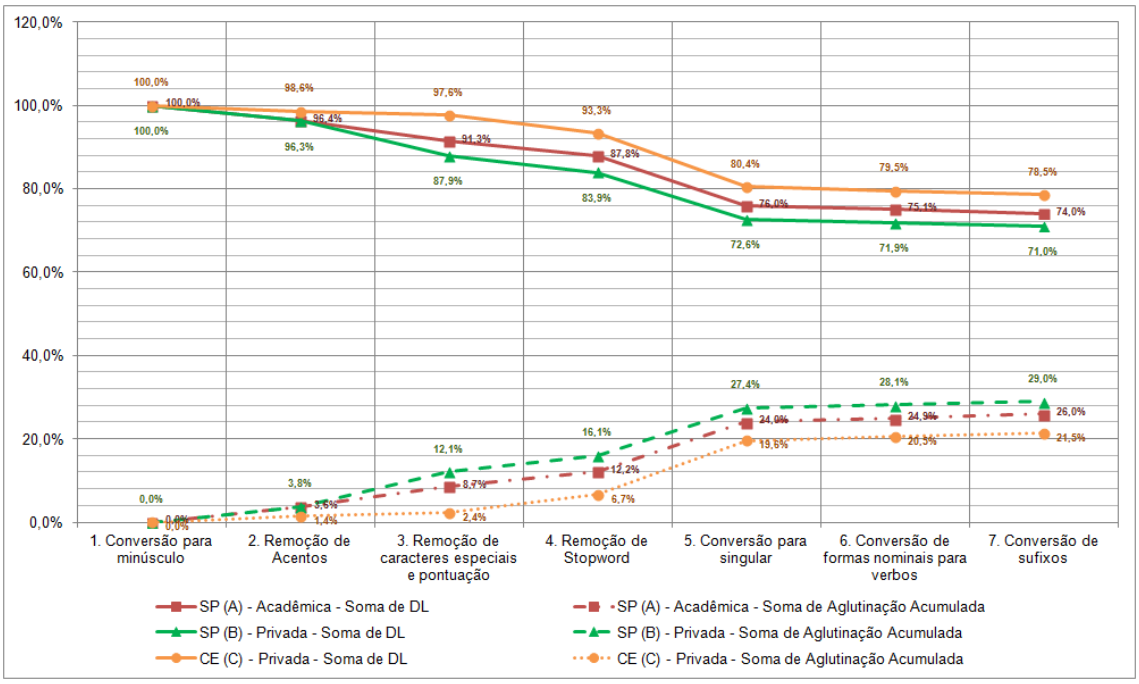


Gráfico 9: Percentual de DL e AL por instituição e tipo de normalização

A subtração entre a AL de um tipo de normalização pela AL do tipo de normalização anterior origina a diferença de aglutinação léxica (dAL) e indica o impacto da normalização (vide gráfico 3).

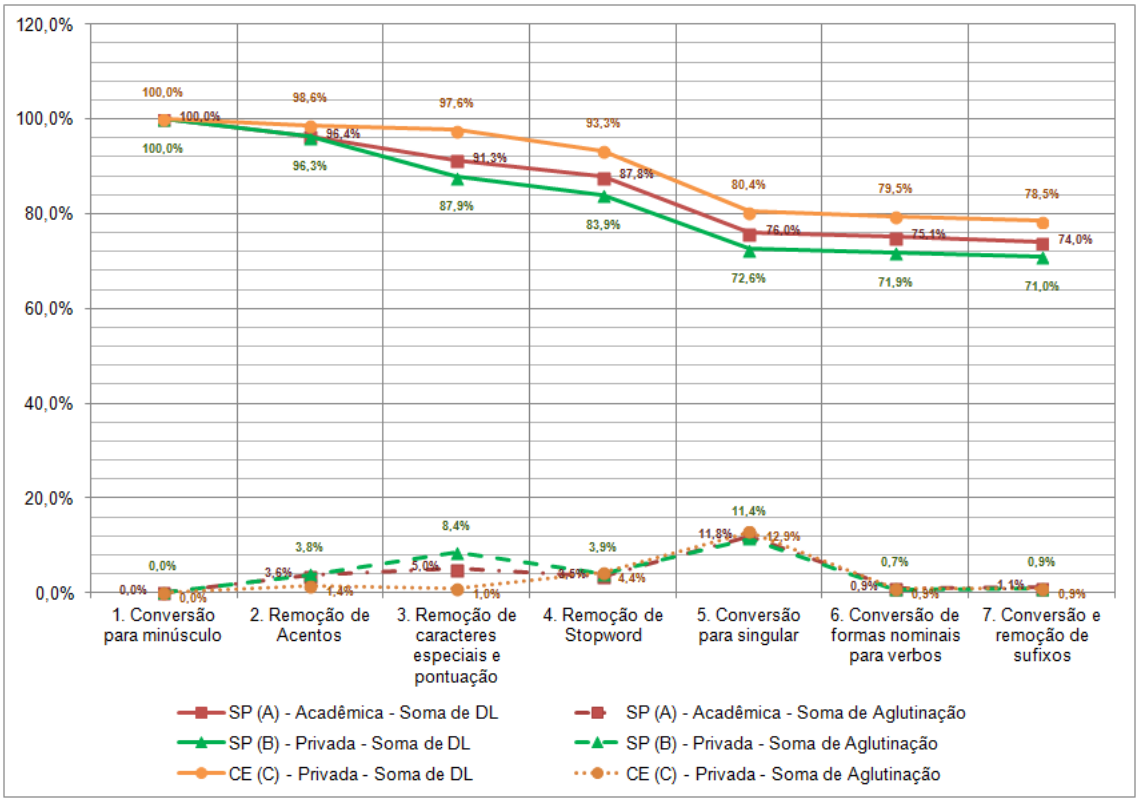


Gráfico 10: DL e dAL por instituição e tipo de normalização

Após aplicação do sétimo e último tipo de normalização a densidade léxica (DL) das instituições A, B e C foi respectivamente de 74,0%, 71,0% e 78,5% e a aglutinação léxica (AL) de 26,0%, 29,0% e 21,5% (vide gráfico 4).

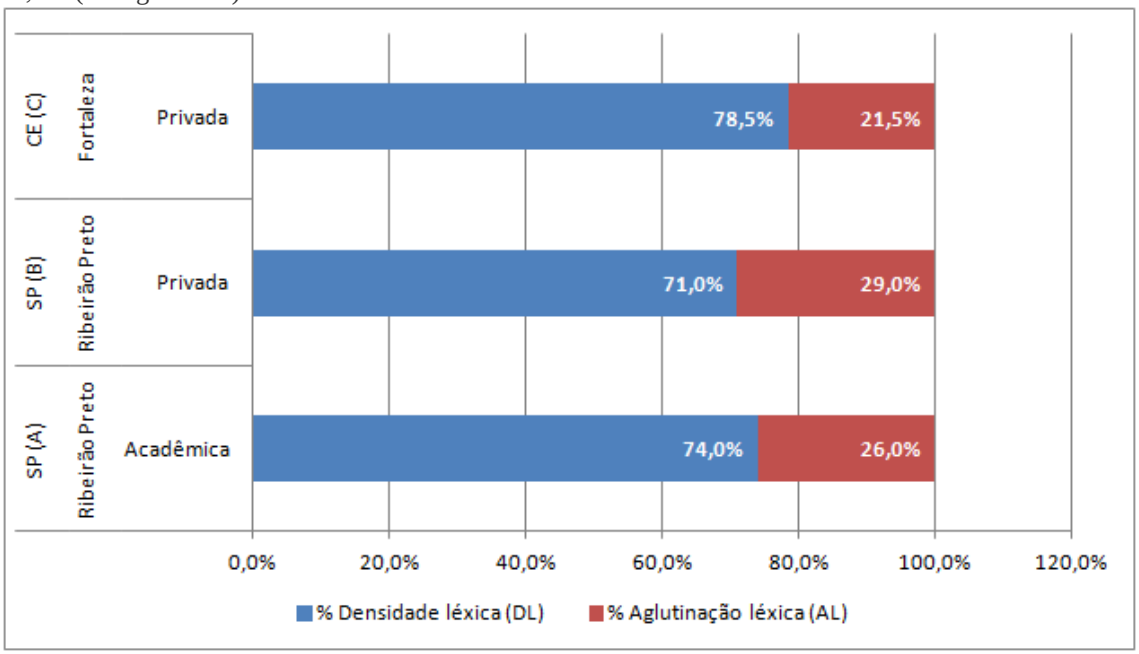


Gráfico 11: Percentual de DL e AL das instituições após a última normalização

Foram identificadas três superestruturas textuais do laudo radiológico, uma para cada instituição, contendo cinco categorias textuais, com variação terminológica em suas denominações, a saber: "Técnica do Exame" (explícita), "Descrição", "Técnica do Exame no primeiro parágrafo da descrição" (implícita), "Conclusão", e "Observação" (vide figura 5).

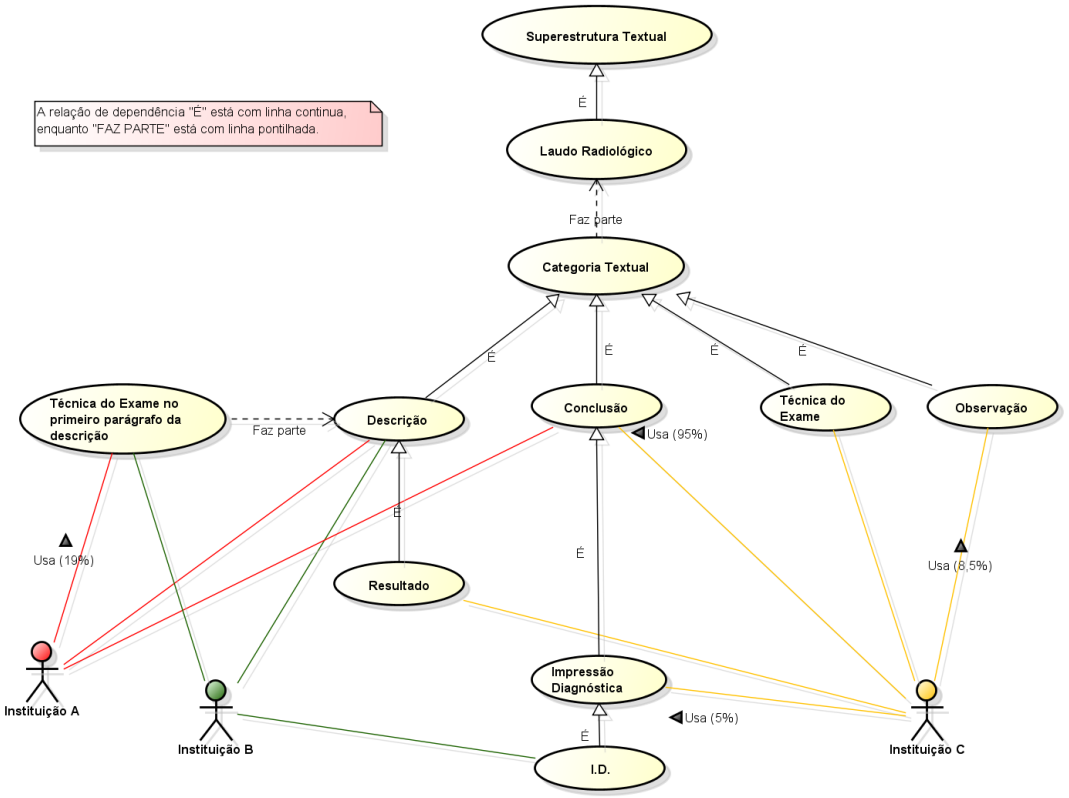


Figura 24: de utilização das superestruturas textuais identificadas nas instituições

A superestrutura textual do laudo radiológico da instituição A é composto pelas categorias textuais "Descrição", "Conclusão" e "Técnica do Exame" no primeiro parágrafo da "Descrição" (implícita), porém, esta última, foi utilizada em 19% dos laudos radiológicos da amostra.

Na instituição B, foram identificadas as categorias textuais "Descrição", "Impressão diagnóstica" e "Técnica do Exame" no primeiro parágrafo da "Descrição" (implícita). A categoria textual "Impressão Diagnóstica" foi identificada pela abreviatura "I.D." e foi utilizada para informar de maneira conclusiva as observações da imagem, assim como, a categoria textual "Conclusão" identificada na instituição A.

No caso da instituição C, foram identificadas as categorias textuais "Resultado", "Conclusão", "Técnica do Exame" e "Observação", todas categorias textuais independentes no texto (explícitas). A categoria textual "Resultado" foi utilizada para descrever os achados da imagem, assim como, a categoria textual "Descrição" identificada na instituição A e B. Observou-se que, para a categoria textual "Conclusão", o termo "Impressão Diagnóstica" foi utilizado em 5% dos laudos da amostra desta instituição, ou seja, ocorreu uma variação terminológica, um caso de sinonímia. A categoria textual "Observação" foi identificada pela abreviatura "OBS:" e ocorreu em 8,5% dos laudos da amostra.

Durante o processo de conceitualização, 4183 termos candidatos foram conceituados originando 727 conceitos únicos em português. As instituições A, B e C compartilham dos conceitos utilizando respectivamente 606, 387 e 431 conceitos (vide tabela 2).

Instituição	Termos Candidatos	Conceitos	Conceitos exclusivos da instituição	% Termos candidatos conceituados	% Termos candidatos conceituados exclusivamente
A	1949	606	185	31%	9%
B	1193	387	38	32%	3%
C	1426	431	67	30%	5%

Tabela 20: Síntese dos conceitos gerados por instituição

Os termos candidatos estão relacionados com os conceitos por 13963 relações onde 6018 relações são usadas na instituição A, 2875 são usadas na instituição B e 5070 são usadas na instituição C. Devido ao compartilhamento dos conceitos entre as instituições a sua distribuição não é uniforme.

Instituição	A		B		C		Considerando todas as instituições	
Relação	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito	Termo candidato com conceito	Conceito com conceito
Associado	7	104	5	82	22	86	34	105
É	1548	250	939	220	998	234	3485	262
Indicado	596	287	367	209	293	193	1256	340
Parte de	3867	80	1564	55	3757	64	9188	82
	6018	721	2875	566	5070	577	13963	789

Tabela 21: Síntese das relações entre termos e conceitos

A primeira instituição que passou pelo processo de conceitualização dos termos candidatos foi a instituição A por possuir a maior quantidade de unidades lexicais normalizadas. Quando os termos candidatos desta instituição foram submetidos de forma automática ao DecS, foram conceituados 22,1% de forma automática e 77,9% manualmente.

A instituição B teve seus termos candidatos submetida de forma automática ao produto terminológico eletrônico (PTE) criado pelo estudo da instituição A e, foram conceituados automaticamente 77,8% dos termos candidatos e 22,2% manualmente.

Por fim, a instituição C, teve seus termos candidatos submetidos de forma automática ao PTE criado pelo estudo da instituição A e B e, foram conceituados automaticamente 83,1% dos termos candidatos e 16,9% manualmente.

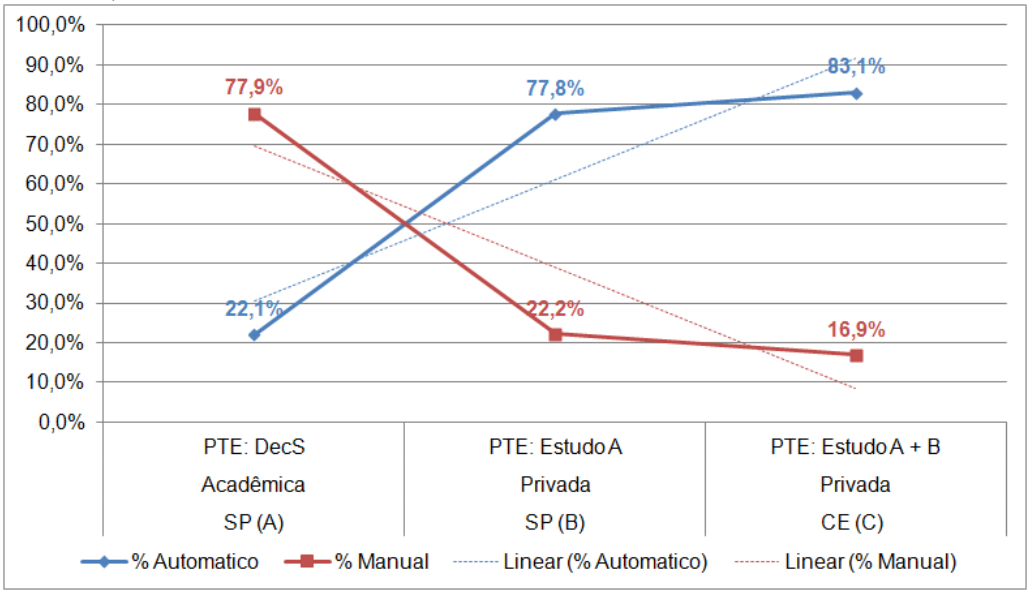


Gráfico 12: Proporção de termos candidatos conceituada manualmente e automaticamente por PTE e instituição

Duas linhas de tendência foram traçadas sobre o gráfico da proporção de termos candidatos conceituada manualmente e automaticamente por PTE e instituição (vide gráfico 7) para evidenciar a tendência de menor necessidade de conceitualização manual quanto maior a quantidade de PTE utilizados, principalmente os extraídos da própria linguagem de especialidade em estudo.

Por se estar propondo uma metodologia, o questionário utilizado é um instrumento exploratório com a finalidade de avaliar a superestrutura e terminologia extraída utilizando a metodologia apresentada. O questionário completo pode ser consultado no apêndice B.

O questionário de avaliação enviado aos 10 (dez, n=10) radiologistas especialistas das instituições participantes desta pesquisa não os identificou e, portanto, todas as respostas são anônimas. Todavia, cada avaliador recebeu um código de identificação sequencial de modo que o primeiro é o A1 e o último A10.

Qualitativo:

Por se estar propondo uma metodologia, o questionário utilizado é um instrumento exploratório com a finalidade de avaliar a superestrutura e terminologia extraída utilizando a metodologia apresentada. O questionário de avaliação enviado aos dez (n=10) radiologistas especialistas das instituições participantes

desta pesquisa não os identificou e, portanto, todas as respostas são anônimas. Todavia, cada avaliador recebeu um código de identificação sequencial de modo que o primeiro é o A1 e o último A10.

Para avaliar as superestruturas identificadas nos laudos do estudo foram apresentadas três superestruturas textuais nomeadas da seguinte forma: “Modelo A”, “Modelo B” e “Modelo C”. A superestrutura do “Modelo A” foi composta por duas categorias textuais denominadas respectivamente “Descrição” e “Conclusão”, onde, dentro da “Descrição” poderia existir uma categoria textual chamada “Observações” a ser preenchida conforme a necessidade. No “Modelo B”, a superestrutura foi constituída por duas categorias textuais explícitas, denominadas respectivamente por “Descrição do laudo” e “Impressão diagnóstica” e, uma categoria textual implícita denominada “Técnica do exame” que é informada no primeiro parágrafo da categoria textual “Descrição do laudo”. Por último, o “Modelo C”, foi constituído de três categorias textuais explicitas denominada respectivamente “Técnica do Exame”, “Descrição” e “Resultado”.

Os modelos apresentados estão baseados nas superestruturas observadas na amostra, porém eles não são exatamente equivalentes nas três instituições, modificando por vezes, a posição das categorias textuais e o termo que descreve cada categoria textual aderente a superestrutura.

Questão 1: “Qual a estrutura de laudo mais adequada?”

Modelo	Quantidade de respostas
A	0
B	6
C	4

Tabela 22: Tabulação das respostas dos avaliadores à questão 1 (n=10)

Nenhum avaliador escolheu o “Modelo A”, conforme observado nas justificativas apresentadas na questão dois, porque não havia no “Modelo A” informação sobre a técnica do exame e o termo “Impressão Diagnóstica” é mais apropriado que os termos “Conclusão” ou “Resultado”.

A maioria dos avaliadores prefere o “Modelo B” porque nele se encontram as categorias textuais essenciais para confecção do laudo radiológico (Prestes-Júnior; Tourinho; Rangel, 2012) e o termo “Impressão Diagnóstica” ao invés de “Conclusão” ou “Resultado”.

A porção de avaliadores que escolheram o “Modelo C”, o fez por considerar mais importante explicitar a categoria textual “Técnica do exame”, do que o termo utilizado para representar a categoria textual conclusiva do laudo radiológico, seja ele, “Impressão diagnóstica”, “Resultado” ou “Conclusão”.

Questão 2: “Qual a estrutura de laudo mais adequada?”

Avaliador	Modelo Escolhido	Resposta
A1	B	Creio que não há uma forma única e correta para relatórios. Em parte a escolha está relacionada a preferências pessoais. Mas há quesitos que são essenciais, como uma parte descritiva e outra na qual o radiologista dá sua opinião baseada em sua experiência e em seu conhecimento de especialista. Eu prefiro que esta última parte se denomine impressão diagnóstica embora não veja como errôneo chama-la de parte conclusiva. Os modelos apresentados me parecem diferir inicialmente mas, na prática podem ser muito semelhantes.
A2	C	Acho que facilita para o médico solicitante visualizar as informações que seriam mais importantes para ele.
A3	B	Sequência mais lógica.
A4	C	Composta da sequencia do tipo de exame; descrição dos achados e a impressão diagnóstica de acordo com a suspeita e achados.
A5	C	Acho que fica mais claro quando se separa a técnica do exame do corpo do laudo. Alem disso acho que o laudo deve ter uma parte descritiva e outra conclusiva.
A6	C	É o modelo que uso.
A7	B	Considero a descrição da técnica do exame importante. Considero que a síntese de achados ou a interpretação do medico Radiologista deve ser denominada impressão diagnostica.
A8	B	Acho a sua estruturação mais adequada, pois apresenta uma impressão diagnóstica baseada na interpretação das imagens e não um "resultado". O modelo A também se aproxima do que acho ideal, mas não se refere à técnica do exame.

A9	B	Ainda na alternativa B, a escolhida, faria observações quando necessário, inclusive sobre a qualidade técnica das imagens.
A10	B	Acredito que o laudo completo deva conter os aspectos técnicos da realização do exame, de preferência em primeiro plano, para já adiantar possíveis problemas ou restrições do exame para a adequada avaliação das imagens. Descrição pormenorizada dos achados e conclusão me parece a sequencia lógica. Creio que o termo impressão diagnostica seja o termo que melhor exprima a avaliação global do exame disposta de forma mais concisa e objetiva dos achados, associada à minha interpretação pessoal no momento de responder aos questionamentos do medico solicitante.

Tabela 23: Tabulação das respostas dos avaliadores à questão 2 (n=10).

Mais da metade dos avaliadores preferem o termo "Impressão Diagnóstica" ao invés do termo "Resultado" contando os que escolheram o "Modelo B" e o avaliador A4 que escolheu o "Modelo C", porém retifica o termo ao escrever "Impressão diagnóstica" ao invés de "Resultado" em sua justificativa da opção por este modelo.

Questão 3: Você sugeriria outra estrutura para o modelo de laudo da coluna lombo sacra? Descreva abaixo. (Se deixar em branco entenderemos que o modelo escolhido é perfeito para elaboração do laudo radiológico da coluna lombossacra).

Avaliador	**Resposta**
A4	Acrescentaria os dados clínicos no início do relatório.
A7	Sugiro acrescentar o espaço para observações, quando necessárias do modelo A.
A8	Adotaria o modelo B. Mas como o modelo A contempla espaço para observações eu também incluiria essa possibilidade no modelo B e utilizaria esse espaço para incluir alterações de outras estruturas que possa ser parcialmente visualizadas no exame e para comparação entre laudos.
A10	Não vejo necessidade de laudo especifico e diferenciado para coluna lombossacra.

Tabela 24: Tabulação das respostas dos avaliadores à questão 3 (n=10).

Menos da metade dos avaliadores responderam a esta questão. Daqueles que responderam, nenhum sugeriu uma nova informação que devesse ser preenchida pelo radiologista ou uma nova superestrutura textual.

Os dados clínicos, sugerido pelo avaliador A4, fazem parte da solicitação do exame e é preenchido pelo medico solicitante. Geralmente o laudo radiológico é escrito acoplado ao sistema de informações radiológicas (RIS, Radiology Information System) que recupera as informações clinicas inseridas pelo medico solicitante e as disponibiliza para o medico executante, no caso, o radiologista. Os avaliadores A7 e A8 sugerem que a categoria textual para informação de observação deva ser incluída no modelo que escolheram e ser preenchida quando necessário. Do ponto de vista do avaliador A10, a superestrutura apresentada é suficientemente abrangente para contemplar o laudo de ressonância magnética da coluna lombossacra.

Questão 4: Os termos apresentados são adequados para a descrição de achados?

Resposta	**Quantidade de respostas**
Concordo plenamente	4
Concordo parcialmente	6
Discordo plenamente	0
Discordo parcialmente	0
Não posso avaliar	0

Tabela 25: Tabulação das respostas dos avaliadores à questão 4 (n=10).

Nenhum avaliador discordou que os termos são adequados e todos puderam avaliar os termos. Todavia, a maioria dos avaliadores discorda da definição de alguns termos como será possível constatar nas respostas da questão número seis.

Questão 5: Qual a quantidade de termos apresentados que é empregada na sua instituição para elaboração do laudo?

Opções	Quantidade de respostas
Nenhum dos termos apresentados	0
1 e 8 termos	0
9 e 15 termos	0
16 e 23 termos	3
24 e 30 termos	7

Tabela 26: Tabulação das respostas dos avaliadores à questão 5 (n=10).

Em consequência ao sorteamento dos termos a serem apresentados aos avaliadores, a proporção de termos entre as instituições não ficou uniforme, sendo distribuída pelas proporções: 100% dos termos existiam na instituição A, 60% dos termos (18 termos) existiam na instituição B e C. O resultado esperado, considerando a distribuição dos termos entre as instituições, seria 40% dos avaliadores tivessem escolhido entre 16 e 23 termos.

Questão 6: Dos termos e conceitos apresentados, existe algum que não corresponde ao significado empregado na sua instituição (clínica)? Qual ou quais?

Avaliador	Resposta
A1	Espondilose.
A3	Campo visual e espondilose.
A5	Hidrossiringomielia: Ha uma diferença entre hidromielia e siringomielia em termos conceituais, porem como não se pode diferencia-las pala RM se conclui como uma junção dos termos, porém o conceito aqui apresentado é o de hidromielia, ou seja, especificamente a dilatação do canal central da medula, enquanto a seringomielia refere-se a qualquer cavidade cística intramedular fora do cana central da medula.
A9	Hidrossiringomielia.

Tabela 27: Tabulação das respostas dos avaliadores à questão 6 (n=10).

Dos trinta termos apresentados quatro, foram considerados não correspondentes ao significado empregado na instituição do avaliador.

Questão 7: Facilitaria a emissão do laudo radiológico a existência de um vocabulário comum entre diferentes instituições.

Opções	Quantidade de respostas
Concordo plenamente	6
Concordo parcialmente	4
Discordo plenamente	0
Discordo parcialmente	0
Não posso avaliar	0

Tabela 28: Tabulação das respostas dos avaliadores à questão 7 (n=10).

Embora, alguns avaliadores não concordem plenamente, a maioria deles está de pleno acordo que um produto terminológico comum entre diversas instituições facilitaria a emissão do laudo radiológico. Nenhum dos avaliadores discordou que existe facilidade em um produto terminológico comum a várias instituições.

Questão 8: Por favor, se preciso, adicione seus comentários.

Avaliador	Resposta
A4	A padronização do laudo apresenta benefícios para o paciente em primeiro lugar, por padronizar também a conduta e assim contribuir para um melhor resultado do tratamento.
A5	Excelente iniciativa. Importantíssimo tentar homogeneizar terminologia em laudos de coluna que são particularmente heterogêneas. Porém senti falta da terminologia usada para alterações em discos intervertebrais como hérnias, abaulamentos protrusões etc., que apesar de existir um consenso publicado ainda e objeto de muita discussão.
A5	Acredito que uma uniformização de laudo e termos seja benéfico para o trafego de informações entre o paciente e diferentes médicos ou então mesmo numa conversa de bar entre o paciente e amigos. No entanto, creio que seja de pouca relevância dentro do

	contexto medico vide que embora exista pluralidade de termos e expressões, o entendimento dos diversos laudos por diferentes médicos se faz sem muitas dificuldades. Talvez seja possível o desenvolvimento de ferramentas que se utilizam de laudos uniformes no intuito de facilitar a confecção de laudos, reduzindo o tempo necessário para sua elaboração. E isso pode realmente funcionar em alguns casos porem eu acredito que cada paciente é um paciente e cada exame é um exame, independente das semelhanças que existam. Sabe-se que os mesmos achados de hérnias discais, por exemplo, podem ter significados clínicos completamente diferentes. O radiologista deve sempre saber relevar os achados dos exames, com base na clinica do paciente fornecida pelo medico solicitante e nos questionamentos do solicitante que motivaram a realização do exame. Isso se dá de forma mais pratica com um texto livre e particularizado para cada exame. Por fim, gostaria de salientar que no meu entendimento, os laudos radiológicos são diálogos entre médicos. Nesse diálogo, o radiologista obrigatoriamente deve ter a liberdade para mudar as estruturas de laudo quando isso se faz necessário, a fim de fazer as suas analises e interpretação do exame prevalecerem de forma mais clara e nítida. A pratica médica, incluindo a radiologia, é também uma espécie de arte, aonde é difícil ser impessoal e robotizado. "Acima de todas as liberdades, dê-me a de saber, de me expressar, de debater com autonomia, de acordo com minha consciência." John Milton.

Tabela 29: Tabulação das respostas dos avaliadores à questão 8 (n=10).

DISCUSSÃO:

O processo de construção da ferramenta computacional *meta-genaralis* foi motivado, principalmente, por não existir uma ferramenta capaz de minerar textos em português e, de forma integrada, permitir a conceituação contextualizada dos termos candidatos em estrutura ontológica e armazenada em RDB. A maior parte dos programas ou mineram textos ou criam ontologias e raramente utilizam RDB que, permite inferências a qualquer momento. Além disso, não foi localizado nenhum programa que considere ou identifique as superestruturas textuais presentes no corpus textual.

Modelar uma estrutura de dados capaz de acomodar a ficha terminológica (Pavel; Nolet, 2002) por idioma e formalizá-la ontologicamente, criar algoritmos em c# para contemplar os arquivos invertidos, conversão textual, anonimização, gravação no RDB, foram desafios técnicos interessantes desta pesquisa.

Há um conceito importante por trás da ferramenta e da metodologia *Meta-generalis*, o conhecimento não está nas unidades lexicais, mas sim, no que elas representam no mundo real. A variação terminológica existe porque cada termo faz referencia a um conceito que pode não existir, estar mais ou menos completo, dependendo da experiência pessoal e do idioma do sujeito que a recebe. Desta forma, a extração de conhecimento não se dá com a extração das unidades lexicais, mas sim, com a sua conceituação em uma estrutura ontológica, preferencialmente computável, que permita a todos os sujeitos da comunidade de especialidade ao qual o termo pertence conhecer a mesma descrição do objeto, ou seja, tenham um conhecimento único sobre o conceito.

Os sete tipos de normalização conseguiram aglutinar o corpus textual composto pelos laudos radiológicos das três instituições estudadas em média 25,5% conservando as unidades lexicais inteligíveis. O primeiro tipo de normalização, conversão para minúsculo, obviamente não aglutinou o texto, porém permitiu ao computador realizar comparações exatas aumentando a velocidade de processamento, pois não havia preocupação com a diferença entre maiúsculos e minúsculos. O tipo de normalização com maior aglutinação entre todas as instituições foi a transformação do plural para o singular, em média 12%. Houve pico de 8% na remoção dos caracteres especiais e pontuação na instituição B porque os textos usavam caracteres como hífen de marcadores. No demais, a normalização teve igual efeito entre as instituições (vide tabela 8 e gráfico 4 e 5).

A superestrutura textual de laudos radiológicos nas três instituições foi semelhante, porém observou-se variação terminológica (sinonímia) entre os nomes das categorias textuais de uma mesma instituição e entre instituições diferentes. Na instituição A e B a categoria textual destinada à descrição dos achados clínicos denominava-se "Descrição", enquanto na instituição C era "Resultado".

A instituição C apresentou maior organização das categorias textuais as deixando explicitas, destacado do texto, independente das outras categorias textuais como a "Técnica do Exame" que na instituição A e B estava implícita, embutida, no primeiro parágrafo da categoria textual "Descrição". Todavia, na instituição C, em 5% dos laudos continham o termo "Impressão Diagnóstica" ao invés de "Conclusão" na categoria textual destinada a este fim (vide figura 20) e, embora a instituição B, assim como a instituição C, utilizava-se da categoria textual nomeada por “Impressão Diagnóstica” ela estava abreviada como “I.D.” caracterizando outro caso de sinonímia.

A instituição A usou a técnica do exame no primeiro parágrafo em 19% dos seus laudos, provavelmente nos casos em que eram necessárias informações adicionais, visto que, o RIS desta instituição

possui pedido de exame eletrônico com superestrutura, contendo categorias textuais com descritores pré-definidos da região, exame e a técnica solicitada.

O processo de conceitualização é o mais trabalhoso para construção do PTE, porque embora exista a conceituação automática os termos candidatos precisavam ser analisados dentro do seu contexto a fim de verificar a compatibilidade do conceito associado e a unidade lexical normalizada. Foram criados conceitos que representavam classes de objetos, por exemplo "vértebra lombar", e os objetos destas classes relacionados a elas, exemplo "L1" é uma "vértebra lombar".

Utilizar PTE específicos da linguagem de especialidade permite uma maior conceituação porque a diferença entre a recuperação automática de conceitos do DecS e o PTE das instituição A e B foi 61% maior. Não ficaram demonstradas grandes diferenças terminológicas entre as instituições, a quantidade de conceitos exclusivos entre elas é em média 6% dos conceitos.

Foram identificadas três superestruturas textuais diferentes. A quantidade de termos candidatos foi de 4183, todos foram conceituados e, originaram 727 conceitos em português.

Analisando a resposta dos avaliadores, considerando as duas primeiras questões, fica evidente que a avaliação é baseada na experiência pessoal de cada um e não em um conhecimento formalizado, porém eles são unanimes no quesito das informações mínimas para o laudo radiológico, corroborando com o exposto por Prestes-Jr & et.al. em 2012 (Prestes-Júnior; Tourinho; Rangel, 2012): técnica do exame, descrição dos achados e impressão diagnóstica. O avaliador A10 foi bastante preciso em descrever a informação contida na categoria textual "Impressão Diagnóstica" ao escrever:

> *"...a avaliação global do exame disposta de forma mais concisa e objetiva dos achados, associada à minha interpretação pessoal no momento de responder aos questionamentos do medico solicitante"* (A10).

Alinhavado com o que escreve Prestes-Jr:

> *"Nos casos em que o imaginologista está convicto de sua interpretação, o laudo pode ser concluído de forma mais direta e objetiva".*

Segundo as respostas dos avaliadores, pode se concluir que uma superestrutura de laudo radiológico desejável seria composta, de forma explicita, as categorias textuais "Técnica do Exame", "Descrição", "Impressão diagnóstica" e "Observação" quando necessário.

Três termos foram considerados inadequados pelos avaliadores em suas respectivas instituições, são eles: "espondilose", "campo visual" e "hidrossiringomielia". Dos três termos "espondilose" foi considerado inadequado por dois avaliadores. Para verificar e entender essas inadequações as referências destes termos na ferramenta *meta-genaralis* e um radiologista especialista foram consultados. O termo "espondilose" foi conceituado pelo DecS e o especialista explica que a confusão está no texto "qualquer parte das vértebras pode ser afetada" porque este termo é usado quando apenas a vértebra é afetada, portanto, neste caso, o DecS trás a descrição correta do conceito e não há equivoco na sua conceitualização. O termo composto "campo visual", conhecido também por FOV do inglês "Field Of View", foi conceituado utilizando o Dicionário Médico Ilustrado Dorland (Dorland, 1999) e, segundo especialista consultado, trata-se de uma confusão de tradução porque o termo se refere à imagem que está sendo estudada. O termo "Hidrossiringomielia" foi conceituado por um artigo científico de Honorato & et. al. em 2004 (Honorato; Borges, 2004) e, o especialista, assim como o avaliador A5, explica que o exame de ressonância magnética não permite distinguir entre hidromielia e siringomielia.

A metodologia criada é constituída por seis passos: 1) coleta dos laudos radiológicos de determinado tipo de exame; 2) decomposição textual; 3) normalização das unidades lexicais; 4) identificação das superestruturas textuais; 5) conceituação dos termos candidatos; 6) avaliação das superestruturas identificadas e da terminologia extraída pelos especialistas na linguagem de especialidade estudada. A aplicação desta metodologia resulta na identificação de superestruturas textuais, resultado semelhante ao que propõe o BI-RADS, conceituação ontológica da terminologia considerando o idioma e o contexto, resultado semelhante ao SNOMED-CT.

Esta pesquisa foi desenvolvida visando propor uma metodologia para a estruturação dos dados contidos nos laudos radiológicos, porque este são, em geral, escritos em linguagem especializada e em texto livre. Por este motivo, embora a estrutura computacional tenha sido construída prevendo a possibilidade de recuperar informações, ou seja, recuperar os documentos mediante a um vetor de consulta, testes neste sentido não foram desenvolvidos e outras pesquisas neste campo seriam de grande importância para avaliar o quanto à ontologia melhora a recuperação das informações no corpus textual estudado. Exportação e importação da ontologia para ferramentas computacionais como Protegé (Deus & Cabral, 2011) também seriam interessantes em pesquisas futuras, a fim de promover a interoperabilidade entre diferentes sistemas.

CONCLUSÃO:

A metodologia utilizada permitiu a estruturação da informação considerando o idioma português e foi capaz de extrair as unidades lexicais mantendo as referencias com as suas respectivas categorias textuais e as conceituando ontologicamente. A superestrutura identificada vai ao encontro do desejo dos avaliadores.

Portanto, é possível estruturar informações contidas em texto livre considerando sua superestrutura textual e a terminologia empregada em português.

REFERÊNCIAS:

ACR. **ACR Breast Imaging and Radiological Data System Birads.** 3rd. ed. Reston, VA, USA.: [s.n.].

ACR, A. C. OF R. **ACR Practice Guideline for Communication of Diagnostic Imaging Findings.**, 2010. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:ACR+PRACTICE+GUIDELINE+FOR+COMMUNICATION+OF+DIAGNOSTIC+IMAGING+FINDINGS#9>. Acesso em: 24 jun. 2013

ADOBE, S. **Lorem and more.**, 2003. Disponível em: <http://www.adobe.com/cfusion/exchange/index.cfm?ev>

AGGARWAL, C. C.; CHENG XIANG ZHAI. An introduction to text mining. *In*: AGGARWAL, C. C.; ZHAI, C. (Eds.). **Mining Text Data**. Boston, MA: Springer US, 2012. p. 1–10.

ALMEIDA, G. M. DE B. FAZER TERMINOLOGIA É FAZER LINGUÍSTICA. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l.] EDIPUCRS, 2010. p. 72.

ALMEIDA, M. B.; SOUZA, R. R. Avaliação do espectro semântico de instrumentos para organização da informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 16, n. 31, 25 maio. 2011.

BARBOSA, F. *et al.* Radiological reports: a comparison between the transmission efficiency of information in free text and in structured reports. **Clinics (São Paulo, Brazil)**, v. 65, n. 1, p. 15–21, jan. 2010.

BELLATRECHE, L.; AIT-AMEUR, Y.; CHAKROUN, C. A design methodology of ontology based database applications. **Logic Journal of IGPL**, v. 19, n. 5, p. 648–665, 20 maio. 2010.

BIREME. **Guia de atualização do DeCS.** [S.l: s.n.]. p. 30

BORST, P.; AKKERMANS, H.; TOP, J. Engineering ontologies. **International Journal of Human-Computer Studies**, v. 46, n. 2-3, p. 365–406, fev. 1997.

BRAGA, T. M. **Uma ferramenta de mineração de texto em bancos de dados de um hospital universitário utilizando decomposições matriciais.** [S.l.] Universidade Federal do Rio de Janeiro, 2011.

CAVALCANTE, R. B.; SILVA, P. C.; FERREIRA, M. N. Sistemas de informação em saúde: Possibilidades e Desafios. **Revista de Enfermagem da UFSM**, v. 1, n. 2, p. 290–299, 2011.

CEUSTERS, W.; SMITH, B. A unified framework for biomedical terminologies and ontologies. **Studies in health technology and informatics**, v. 160, n. Pt 2, p. 1050–1054, 2010.

CIMIANO, P.; REYLE, U.; ŠARIĆ, J. Ontology-driven discourse analysis for information extraction. **Data & Knowledge Engineering**, v. 55, n. 1, p. 59–83, out. 2005.

CONRADO, M. DA S.; GUTIÉRREZ, V. A. L.; REZENDE, S. O. Evaluation of Normalization Techniques in Text Classification for Portuguese. *In*: MURGANTE, B. *et al.* (Eds.). **Computational Science and Its Applications – ICCSA 2012**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 7335p. 618–630.

CORRÊA, R.; MIRANDA, D. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **A.toZ: novas práticas em informação e conhecimento**, v. 1, n. 1, p. 11–22, 2011.

DBDESIGNER. **DBDesigner Fork**. Disponível em: <http:sourceforge.net/projects/dbdesigner-fork/>. Acesso em: 18 jun. 2012.

DENA, C.; DOUGLASS, J.; MARINO, M. **Benchmark fiction: A framework for comparative new media studies.**In Proceedings of the 6th Digital Arts and Culture Conference (DAC 2005). **Anais**...2005Disponível em: <http://www.christydena.com/wp-content/uploads/2007/11/2005-dac-benchmark-fiction-final-print-submission1.pdf>. Acesso em: 24 jun. 2013

DEUS, W. DE; CABRAL, E. Um estudo de ontologias como ferramenta para a modelagem do processo de comunicação entre agentes. **Revista Crase. edu**, v. 01, n. 1, p. 16–37, 2011.

DORLAND, W. A. N. **Dicionário Médico Ilustrado Dorland**. 28ª. ed. Barueri-SP: Editora Manole, 1999. p. 2217

DUQUE, J. L. **Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico.** [S.l: s.n.].

DUQUE, J.; MATOS, P.; CIFERRI, C. DE A. Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico. v. d, 2012.

ELDER, J.; MINER, G.; NISBERT, B. **Practical text mining and statistical analysis for non-structured text data applications.** Oxford: Elsevier Inc., 2012. p. 1053

FAULSTICH, E. Socioterminologia: mais que um método de pesquisa, uma disciplina. **Ciência da Informação**, v. 24, 1995.

FENELON, S. Aspectos ético-legais em Imaginologia. **Radiologia Brasileira**, v. 36, n. 1, 2003.

FRANCELIN, M. M.; KOBASHI, N. Y. Concepções sobre o conceito na organização da informação e do conhecimento. **Ciência da Informação**, v. 40, n. 2, p. 207–228, 2011.

FREIRE, P. DE S. *et al.* Memória organizacional e seu papel na gestão do conhecimento. **Ciências da Administração**, v. 14, n. 33, p. 41–51, 2012.

HONORATO, D.; BORGES, W. Syringohydromyelia associated to therapeutic procedures for severe forms of neurocysticercoses: case report. **Arq Neuropsiquiatr**, v. 62, n. 3-b, p. 885–888, 2004.

HUSSAIN, M. W.; AFZAL, M. T.; WAQAS, M. Finding Semantic Relationship among Associated Medical Terms. **Int. J. Emerg. Sci**, v. 2, n. 2, p. 300–309, 2012.

IHTSD, I. H. T. S. D. **SNOMED CT User Guide.** Disponível em: <www.snomed.org/ug.pdf>. Acesso em: 28 dez. 2012.

KAHN, C. E. Multilingual retrieval of radiology images. **Radiographics: a review publication of the Radiological Society of North America, Inc**, v. 29, n. 1, p. 23–9, 2009.

KONG, Z.; ZHANG, K. **Summarizing Keyword Search in Relational Database.**Third International Conference on Intelligent Human-Machine Systems and Cybernetics. **Anais**...IEEE, ago. 2011Disponível

em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6038237>. Acesso em: 24 jun. 2013

KRAUSE, D. **Tópicos em Ontologia Analítica.** [S.l: s.n.].

LANGLOTZ, C. P. RadLex: a new method for indexing online educational materials. **Radiographics: a review publication of the Radiological Society of North America, Inc**, v. 26, n. 6, p. 1595–7, 2006.

MACHADO, A. DA S. *et al.* Reestruturação produtiva no setor de saúde: estudo de caso em um hospital de Fortaleza-CE. **Revista de Administração Hospitalar e Inovação em Saude**, p. 65–74, 2012.

MACIEL, A. M. B. Linguagens especializadas e terminologia: o passado projetando o futuro. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l.] EDIPUCRS, 2010. p. 28.

MACIEL-LIMA, S.; RASIA, J. As Imagens não Falam por si Mesmas: um estudo sociológico da equipe do transplante hepático Images. **Saúde Soc**, v. 21, n. 1, p. 42–51, 2012.

MARTINS, V. F. **Avaliação de usabilidade para sistemas de transcrição automática de laudos em radiologia.** [S.l: s.n.].

MHAMDI, F.; ELLOUMI, M. A new survey on knowledge discovery and data mining. **2008 Second International Conference on Research Challenges in Information Science**, p. 427–432, jun. 2008.

MHIRI, S.; DESPRES, S. Ontology usability via a visualization tool for the semantic indexing of medical reports (DICOM SR). *In*: HOLZINGER, A. (Ed.). **HCI and Usability for Medicine and Health Care**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4799p. 409–414.

MICHAELIS. **MICHAELIS, Moderno Dicionário da Língua Portuguesa.** Disponível em: <http://michaelis.uol.com.br/moderno/portugues/index. php?>.

NARDON, F. B. **Compartilhamento de conhecimento em saúde utilizando ontologias e bancos de dados dedutivos.** [S.l.] Escola Politécnica da Universidade de São Paulo, 2003.

NETTER, F. H. **Atlas de Anatomia Humana**. 2ª. ed. [S.l.] Artmed, 2000. p. 142

OLIVEIRA, A. DE; MOTTA, R. DA; CUNHA, G. Mineração de textos: uma experiência usando TMSK e RIKTEXT. p. 1–11, 2011.

OLIVEIRA, A. M. P. P. DE; ISQUERDO, A. N.; ALVES, I. M. **As ciências do léxico: lexicologia, lexicografia, terminologia.** [S.l.] Editora Ufms, 2007. v. 3

PADILHA, T. P. P.; LACERDA, A. N. **Text Recognition for Concept Maps Construction in Collaborative Environments.**Brazilian Symposium on Collaborative Systems. **Anais**...IEEE, out. 2012Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6391724>. Acesso em: 12 jun. 2013

PASSINI, M. L. C. **Mineração de textos para organização de documentos em centrais de atendimento.** [S.l: s.n.].

PASSMAN, L. J. *et al.* SISMAMA--implementation of an information system for breast cancer early detection programs in Brazil. **Breast (Edinburgh, Scotland)**, v. 20 Suppl 2, p. S35–9, abr. 2011.

PAVEL, S.; NOLET, D. **Manual de terminología.** [S.l.] Ministério de Obras Públicas e Governamentais do Canadá, 2002.

PINTO, A.; BRUNESE, L. Spectrum of diagnostic errors in radiology. **World journal of radiology**, v. 2, n. 10, p. 377–83, 28 out. 2010.

PINTO, T. L. **Aspectos da terminologia da economia internacional.** [S.l: s.n.].

PRESTES-JÚNIOR, L. C. L.; TOURINHO, E. K.; RANGEL, M. Análise médico-legal das demandas judiciais em imaginologia. **Radiologia Brasileira**, v. 45, n. 2, p. 98–100, 2012.

QUANTIN, C. *et al.* Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm. **Revue d'Épidémiologie et de Santé Publique**, v. 57, n. 1, p. e1–e6, fev. 2009.

RAHMAN, M. Information retrieval with text mining for Decision Support System. **Bangladesh Journal of Scientific Research**, v. 24, n. 2, p. 117–126, 2011.

REIS, A. DE O. As interfaces das disciplinas do Léxico. **Dialogia**, v. 9, n. 2, p. 173–179, 2010.

REZENDE, S.; MARCACINI, R.; MOURA, M. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, v. 7, p. 7–21, 2011.

ROBERTO, J.; MARTÍ, M.; LLORENTE, M. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. **Procesamiento de Lenguaje …**, v. 48, p. 97–104, 2012.

SCHAEFER, F. Proteção de dados de saúde como direito fundamental. **Cadernos da Escola de Direito e Relações Internacionais**, v. 17, p. 1–19, 2012.

SILVA-FILHO, S. C. DA. A VARIAÇÃO E A RELAÇÃO CONCEITO/TERMO: UMA QUESTÃO DE PONTO DE VISTA. **Trauma**, v. 6, n. 12, p. 75–86, 2010.

UMLS. **Unified Medical Language System**. Disponível em: <https://uts.nlm.nih.gov/home.html>. Acesso em: 19 jun. 2013.

VASILÉVSKI, V. Tratamento dos sufixos modo-temporais na depreensão automática da morfologia dos verbos do português. **Linguamática**, v. 3, n. 2, p. 107–118, 2011.

VIEIRA, J. DE L.; CORREA, R. VISUALIZAÇÃO DA INFORMAÇÃO NA CONSTRUÇÃO DE INTERFACES AMIGÁVEIS PARA SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO. **Encontros Bibli**, v. 16, n. 32, p. 73–93, 2011.

VIEIRA, M. C. T. **Leitura no ensino superior: letramentos em pauta.**XV ENDIPE. **Anais**...2010

VIEIRA, R.; LOPES, L. PROCESSAMENTO DE LINGUAGEM NATURAL E O TRATAMENTO COMPUTACIONAL DE LINGUAGENS CIENTÍFICAS. *In*: **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. [S.l: s.n.]. .

VILLAVICÊNCIO, A.; RAMISCH, C. Chutando o balde ou batendo as botas? Processamento de linguagem natural e expressões multipalavra na linguagem científica. *In*: PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. (Eds.). **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. Porto Alegre: Edipucrs, 2010. p. 29–49.

ZERBINATTI, L. **Extração de Conhecimento em Laudos de Radiologia Torácica na Língua Portuguesa.** [S.l.] Escola Politécnica da Universidade de São Paulo, 2010.

ZILIO, L. TERMO E VALOR LINGUÍSTICO: UMA ABORDAGEM ENSAÍSTICA. **Cadernos do IL**, v. 42, p. 119–127, 2012.