PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE

FARUK MUSTAFA ZAHRA

PORONTO - FERRAMENTA PARA CONSTRUÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS EM PORTUGUÊS

CURITIBA

2009

FARUK MUSTAFA ZAHRA

PORONTO - FERRAMENTA PARA CONSTRUÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS EM PORTUGUÊS

Dissertação apresentada ao Curso de Pós Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde, como parte dos requisitos para a obtenção do título Mestre em Tecnologia em Saúde.

Área de Concentração: Informática em Saúde

Orientadora: Prof[a]. Dr[a]. Andreia Malucelli

CURITIBA

2009

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

Z19p
2009

Zahra, Faruk Mustafa
Poronto : ferramenta para construção semiautomática de ontologias em português / Faruk Mustafa Zahra ; orientadora, Andreia Malucelli. – 2009.
92 p. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2009
71-77

1. Inteligência artificial. 2. Ontologia. 3. Informática na medicina. I. Mulucelli, Andreia. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Tecnologia em Saúde. III. Título.

CDD 20. ed. – 610.28

Pontifícia Universidade Católica do Paraná
Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Tecnologia em Saúde

ATA DA SESSÃO PÚBLICA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
DO PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE
DA PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

DEFESA DE DISSERTAÇÃO Nº 116

Aos 17 dias de dezembro de 2009 realizou-se a sessão pública de defesa da dissertação "**PORONTO - Ferramenta para Construção Semiautomática de Ontologias em Português**", apresentada por **Faruk Mustafa Zahra** como requisito parcial para a obtenção do título de Mestre em Tecnologia em Saúde – Área de Concentração – **Informática em Saúde** perante uma Banca Examinadora composta pelos seguintes membros:

Profª. Drª. Andreia Malucelli, PUCPR (Orientadora)	assinatura	APROVADO (parecer (aprov/ reprov.))
Profª. Drª. Marcia Regina Cubas, PUCPR		APROVADO
Prof. Dr. Emerson Cabrera Paraiso, PUCPR/PPGIa		APROVADO
Prof. Dr. César Augusto Tacla, UTFPR		aprovado

Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado APROVADO (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do Programa.

Prof. Dr. Munir Antonio Gariba,
Diretor do PPGTS PUCPR

Pós-Graduação Tecnologia em Saúde - CCBS PUCPR

Rua Imaculada Conceição, 1155 Prado Velho CEP 80215 901 Caixa Postal 17315 CEP 80242 980
Tel./Fax (41) 3271 1657 E-mail ppgts@pucpr.br www.pucpr.br Curitiba Paraná Brasil

Dedico esta dissertação ao meu Pai,
mesmo sem o título, é o nosso Mestre.

AGRADECIMENTO

Obrigado Allah por me dar esta chance tão valiosa e estar sempre Presente.

À minha amada mãe por sempre acreditar em mim, mesmo com todas as professoras e diretoras dizendo o contrário, valeu a pena mãe ter investido em uma criança tão "problemática".

Ao meu querido e admirável pai, obrigado por ter abdicado de uma carreira acadêmica brilhante por nós.

À minha amada esposa, sem você eu não teria forças para continuar este trabalho, seu carinho e suas doces palavras de incentivo fizeram toda a diferença!

À minha querida orientadora, somente entendi o sentido da palavra "professor" após te conhecer, obrigado pela paciência, dedicação e principalmente o amor pela docência.

À professora Marcia Regina Cubas e seu orientando Paulo Cesar Zimmermann Felchner, obrigado pelo tempo despendido na colaboração deste trabalho.

RESUMO

Aprendizagem de ontologias (*ontology learning*) é definida como um conjunto de métodos e técnicas para a construção semiautomática de uma nova ontologia ou para o enriquecimento de uma já existente. Para construir uma ontologia de maneira semiautomática é necessário a automatização do processo de aquisição de conhecimento, e dentre as abordagens sugeridas, a construção de ontologias a partir de texto foi a abordagem utilizada. O objetivo desta pesquisa foi desenvolver uma ferramenta semiautomática para construção de ontologias a partir de textos em português na área da saúde. A ferramenta proposta foi desenvolvida tendo como base a ferramenta para construção semiautomática de ontologias a partir de textos em inglês Text-to-Onto; as medidas de frequência, *tfidf* e entropia; e a ferramenta de anotação linguística TreeTagger. Foram realizados três experimentos para validar a ferramenta, sendo os experimentos nas áreas de domínio do câncer de mama; de eventos adversos pós-vacinação e de inventário vocabular em enfermagem. Os termos extraídos foram validados por especialistas nas respectivas áreas. Concluiu-se que a quantidade de termos extraídos pela ferramenta é dependente da poda realizada pelos filtros que eliminam termos simples e compostos de acordo com a frequência estabelecida; a alta frequência com que um termo aparece no *corpus* não significa que ele será selecionado, devido à quantidade de ruído no *corpus*; é importante incluir outros tipos de filtros, para refinar a extração dos termos; o uso dos Descritores em Ciência da Saúde (DeCS) como medida para extração pode melhorar os resultados obtidos. Este trabalho é uma contribuição para o processo de aprendizagem de ontologias, fornecendo uma ferramenta com funcionalidades até então não propostas: não há necessidade do usuário realizar previamente a anotação linguística do *corpus*; a ferramenta está disponível em ambiente *Web*; há inserção de sinônimos; e identificação de termos DeCS.

Palavras-Chaves: Inteligência Artificial. Vocabulário Controlado. Ontologia. Aprendizagem de Ontologias.

ABSTRACT

Ontology Learning is defined as a set of methods and techniques for the semi-automatic construction of a new ontology or to enrich an existing one. To build an ontology in a semi-automatic way it is necessary to automate the process of acquiring knowledge. There are different approaches and in this research it was used construction of ontologies from text. This research aims to develop a tool for semi-automatic ontology construction from texts in Portuguese, in the health domain. The proposed tool was developed based on Text-to-Onto, which is a tool for semi automatic ontology construction from texts in English; frequency, tfidf, and entropy measures; and the tool for linguistic annotation, TreeTagger. Three experiments were conducted to validate the tool, in the domains of breast cancer, adverse events post vaccination, and vocabulary inventory in nursing. The extracted terms were validated by experts in their specific domains. It is possible to conclude that the number of extracted terms by the tool is dependent on the pruning carried out by the filters that eliminate simple terms and compound terms according to the established frequency; the high frequency of a term in the corpus does not mean it will be selected because of the amount of "noise" in the corpus; it is important to include other types of filters, to refine the terms extraction; the use of Descriptors in Health Science (DeCS) as a measure for extraction may improve the results. This research is a significant contribution to ontology learning process, providing a tool with features not previously proposed: it is not necessary for the user to perform the linguistic annotation of corpus; the tool is available on the Web platform; there is the identification of synonyms for terms; and the identification of terms in the DeCS descriptor.

Key Words: Artificial Intelligence. Controlled Vocabulary. Ontology. Ontology Learning.

LISTA DE FIGURAS

FIGURA 1: EXEMPLO DA ABORDAGEM BASEADA EM TERMOS COMPOSTOS..... 30
FIGURA 2: ABORDAGEM PROPOSTA POR BASÉGIO. 33
FIGURA 3: EXEMPLO DO FORMATO XCES 34
FIGURA 4: INTERFACE GRÁFICA DO TERMINAE 38
FIGURA 5: TERMOS EXTRAÍDOS NO TEXT-TO-ONTO 39
FIGURA 6: ARQUIVO DE ENTRADA PARA O TREETAGGER 42
FIGURA 7: ARQUIVO DE SAÍDA DO TREETAGGER 42
FIGURA 8: VISL EM MODO ONLINE 43
FIGURA 9: EXEMPLO DE INSTRUMENTO ENVIADO AO ESPECIALISTA PARA VALIDAÇÃO 48
FIGURA 10: PROTÓTIPO 50
FIGURA 11: PROCESSO DE CRIAÇÃO DO *CORPUS* DO PROTÓTIPO 51
FIGURA 12: PROCESSO DA CRIAÇÃO DA ONTOLOGIA DO PROTÓTIPO 51
FIGURA 13: IMPORTAR *CORPUS* 52
FIGURA 14: EXTRAÇÃO DE TERMOS SIMPLES NO PROTÓTIPO 52
FIGURA 15: EXTRAÇÃO DE TERMOS COMPOSTOS NO PROTÓTIPO 53
FIGURA 16: EXPORTAR PARA XLS 54
FIGURA 17: PÁGINA DE ENTRADA 55
FIGURA 18: PROCESSO DE CRIAÇÃO DO *CORPUS* DA FERRAMENTA 56
FIGURA 19: ENVIO DOS ARTIGOS PARA PROCESSAMENTO 56
FIGURA 20: PROCESSO DA CRIAÇÃO DA ONTOLOGIA DA FERRAMENTA 57
FIGURA 21: EXTRAÇÃO DE TERMOS SIMPLES NA FERRAMENTA 59
FIGURA 22: EXTRAÇÃO DE TERMOS COMPOSTOS NA FERRAMENTA 59
FIGURA 23: SINÔNIMOS 59
FIGURA 24: LISTA DOS TERMOS PARA SELEÇÃO 60
FIGURA 25: MENU GERAR ONTOLOGIA 60
FIGURA 26: TAXONOMIA GERADA 61
FIGURA 27: ARQUIVO XLS GERADO 62
FIGURA 28: ARQUIVO OWL GERADO 62
FIGURA 29: ONTOLOGIA IMPORTADA PARA O PROTÉGÉ 63

LISTA DE QUADROS

QUADRO 1: ONTOLOGIAS NA ÁREA DE SAÚDE 21
QUADRO 2: PADRÃO DE HEARST 26
QUADRO 3: PADRÃO DE MORIN/JACQUEMIN 28
QUADRO 4: COMPARATIVO ENTRE FERRAMENTAS PARA APRENDIZAGEM DE ONTOLOGIAS PARA LÍNGUAS ESTRANGEIRAS 32
QUADRO 5: TERMOS EXTRAÍDOS PELO KEA 37
QUADRO 6: AVALIAÇÃO FINAL DAS FERRAMENTAS PARA APRENDIZAGEM DE ONTOLOGIA. 40
QUADRO 7: AVALIAÇÃO DAS FERRAMENTAS DE ANOTAÇÃO LINGUÍSTICA. 44
QUADRO 8: REGRAS DE IDENTIFICAÇÃO DE TERMOS COMPOSTOS NO PROTÓTIPO 53
QUADRO 9: REGRAS DE IDENTIFICAÇÃO DE TERMOS COMPOSTOS NA FERRAMENTA 58
QUADRO 10: SÍNTESE DOS TERMOS REPETIDOS, TERMOS ÚNICOS E TERMOS EXTRAÍDOS 63
QUADRO 11: PARÂMETROS PARA OS FILTROS UTILIZADOS 64
QUADRO 12: TERMOS NÃO SELECIONADOS PELOS ESPECIALSITAS NO EXPERIMENTO 2 65
QUADRO 13: TERMOS NÃO SELECIONADOS PELOS ESPECIALISTAS NO EXPERIMENTO 1 66
QUADRO 14: QUANTIDADE DOS TERMOS SELECIONADOS QUE SÃO DECS 67
QUADRO 15: DISTRIBUIÇÃO DOS TERMOS SELECIONADOS LEVANDO EM CONSIDERAÇÃO A MEDIDA DE FREQUÊNCIA 67
QUADRO 16: PORCENTAGEM DE ACERTO EM CADA REGRA DE EXTRAÇÃO DE TERMOS COMPOSTOS NO EXPERIMENTO 1 69
QUADRO 17: PORCENTAGEM DE ACERTO EM CADA REGRA DE EXTRAÇÃO DE TERMOS COMPOSTOS NO EXPERIMENTO 2 69
QUADRO 18: PORCENTAGEM DE ACERTO EM CADA REGRA DE EXTRAÇÃO DE TERMOS COMPOSTOS NO EXPERIMENTO 3 69
QUADRO 19: RESULTADO NO EXPERIMENTO NA ÁREA DE CÂNCER DE MAMA 85
QUADRO 20: RESULTADO DO EXPERIMENTO DA ÁREA DE EVENTOS ADVERSOS PÓS-VACINAÇÃO 90
QUADRO 21: RESULTADO DO EXPERIMENTO DA ÁREA DE INVENTARIO VOCABULAR 94

LISTA DE GRÁFICOS

GRÁFICO 1: COMPARAÇÃO DAS PORCENTAGENS DE TERMOS EXTRAÍDOS CORRETAMENTE 64

GRÁFICO 2: TOTAL DE TERMOS SELECIONADOS PELOS TRÊS ESPECIALISTAS DO EXPERIMENTO 2 65

GRÁFICO 3: PORCENTAGEM DE TERMOS SELECIONADOS PELOS ESPECIALISTAS QUE APARECEM COM MENOR FREQUÊNCIA NO CORPUS 68

GRÁFICO 4: TERMOS COMPOSTOS EXTRAÍDOS E SELECIONADOS 68

GRÁFICO 5: COMPARAÇÃO DOS TERMOS COMPOSTOS SELECIONADOS POR CADA REGRA NOS TRÊS EXPERIMENTOS 70

LISTA DE ABREVIATURAS E SIGLAS

AMD	Advanced Micro Devices
BIREME	Biblioteca Virtual em Saúde
CIPE®	Classificação Internacional para a Prática de Enfermagem
CIPESC®	Classificação Internacional para a Prática de Enfermagem em Saúde Coletiva
DeCS	Descritores em ciência da saúde
FR	Frequência relativa
GO	Gene Ontology
GB	Gigabyte
GHz	Gigahertz
IA	Inteligência Artificial
ICNP®	International Classification for Nursing Practice
IDE	Integrated Development Environment
IMGT	ImMunoGeneTics
INCA	Instituto Nacional do Câncer
JSF	Java Server Faces
KB	Quilobyte
LAEL	Linguística Aplicada e Estudos da Linguagem
PUC-SP	Pontifícia Universidade Católica de São Paulo
MHC	Major Histocompatibility Complex
NP	Noun Phrase
OBO	Open Biomedical Ontologies
OWL	Web Ontology Language

PDF	Portable Document Format
PLN	Processamento de linguagem natural
RAM	Random Access Memory
RDF	Resource Description Framework
SNOMED	Systemized Nomenclature of MEDicine
TF-IDF	Term frequency-inverse document frequency
UMLS	Unified Medical Language System
XLS	Formato utilizado pelo Microsoft Excel
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO 14

1.1 OBJETIVO 16

1.1.1 Objetivo Geral 16

1.1.2 Objetivos Específicos 16

1.2 ESTRUTURA DO DOCUMENTO 16

2 REVISÃO DE LITERATURA 17

2.1 ONTOLOGIAS 17

2.1.1 Tipos de ontologias 19

2.1.2 Ontologias na área de saúde 20

2.2 APRENDIZAGEM DE ONTOLOGIAS 21

2.2.1 Aprendizagem de Ontologias a partir de textos 22

2.2.2 Métodos para a extração de termos 24

2.2.3 Métodos para geração taxonômica 26

2.3 FERRAMENTAS PARA CONSTRUÇÃO DE ONTOLOGIAS 30

2.3.1 Ferramentas para línguas estrangeiras 30

2.3.2 Ferramentas para língua portuguesa 32

3 MÉTODO 36

3.1 IDENTIFICAÇÃO DE FERRAMENTAS PARA APRENDIZAGEM DE ONTOLOGIA 36

3.2 IDENTIFICAÇÃO DE FERRAMENTAS DE ANOTAÇÃO LINGUÍSTICA 40

3.3 ELABORAÇÃO DO PROTÓTIPO 44

3.4 AVALIAÇÃO DOS RESULTADOS DO PROTÓTIPO 45

3.5 ELABORAÇÃO DA VERSÃO FINAL DA FERRAMENTA 45

3.6 VALIDAÇÃO DOS TERMOS EXTRAÍDOS PELA FERRAMENTA 47

3.6.1 Validação na área de câncer de mama 47

3.6.2 Validação na área de eventos adversos pós-vacinação 48
3.6.3 Validação na área de inventário vocabular de enfermagem 49
4 RESULTADOS E DISCUSSÃO 50
4.1 PROTÓTIPO 50
4.2 VALIDAÇÃO DO PROTÓTIPO..... 54
4.3 VERSÃO FINAL DA FERRAMENTA 55
4.4 VALIDAÇÃO DOS TERMOS EXTRAÍDOS PELA FERRAMENTA..... 63
5 CONCLUSÕES..... 71
REFERÊNCIAS 73
APÊNDICE I..... 80
APÊNDICE II 86
APÊNDICE III 91

1 INTRODUÇÃO

Muitas pesquisas em Inteligência Artificial (IA) têm sido dedicadas ao desenvolvimento de sistemas computacionais que incorporam conhecimento sobre determinado domínio, permitindo inferências, raciocínios e tomada de decisões. Estes sistemas mantêm uma representação explícita e simbólica do conhecimento. Tal representação tem a vantagem de estar separada dos aspectos procedurais relacionados à aplicação, podendo desta forma ser reusado por outros sistemas. Para realizar esta tarefa se faz necessário organizar o conhecimento de maneira formal e disponibilizá-lo em uma linguagem padrão para que possa ser compartilhado, pois computadores são, essencialmente, máquinas processadoras de símbolos e precisam de instruções claras sobre como manipular estes símbolos de uma maneira significativa (CIMIANO, 2005).

Neste contexto, as ontologias podem ser utilizadas para representar o domínio em questão, pois permitem representar vocabulários formais que descrevem as premissas básicas de um determinado domínio (FREITAS; SCHULZ, 2009).

A definição de ontologia mais amplamente difundida e utilizada diz que "ontologia é uma especificação explícita de uma conceitualização" (GRUBER, 1993). Esta definição foi complementada, posteriormente, para "ontologia é uma especificação formal de uma conceitualização compartilhada" (BORST, 1997).

Processar informações usando ontologias, que proveem excelente contexto para o entendimento das informações, tanto para usuários humanos quanto para agentes de *software*, vem se tornando uma tendência em várias áreas e tipos de aplicações. Porém, entre as áreas mais comuns de aplicação de ontologia estão a medicina e a biologia. A principal razão para isso pode ser a complexidade do conhecimento médico e biológico tornar difícil a confecção de sistemas tradicionais, pois para assistir tarefas médicas, os sistemas precisam de muito conhecimento e capacidade de inferência (MUSEN, 2002; FREITAS; SCHULZ, 2009).

Pode-se citar alguns exemplos de ontologias na área de saúde, desenvolvidas na língua inglesa: *Unified Medical Language System* (UMLS, 2009), *Open Biomedical Ontologies* (OBO, 2009), *Systemized Nomenclature of MEDicine* (SNOMED, 2009) e *Gene Ontology* (GO, 2009). Para o português, alguns esforços estão sendo realizados para construção de ontologias em saúde, dentre os quais: Ontologia para a Classificação Internacional para a Prática de Enfermagem em Saúde Coletiva (CIPESC®) (SILVA; MALUCELLI; CUBAS,

2008), Ontologia do domínio de Genoma e Transcriptoma (CAMPOS, 2007) e Ontologia na área de Histocompatibilidade no Transplante de Órgãos (ALVES; MALUCELLI, 2008).

Há muito interesse no desenvolvimento de ontologias na área da saúde, porém, há pouco trabalho desenvolvido e pouca utilização de ontologias em larga escala. Algumas das razões para isso são o tempo e custo associados neste desenvolvimento. Para a criação de uma ontologia é necessário que um especialista da área a ser construída a ontologia transfira o conhecimento consensual de sua área para a ontologia, sendo que este processo é extremamente custoso. Além disso, as ontologias devem ser compartilhadas por um grupo de pessoas ou por uma comunidade, sendo o processo de construção dificultado por envolver diferentes pessoas com pontos de vistas, muitas vezes, divergentes (CIMIANO, 2005).

Este fato levou ao surgimento de uma nova área de pesquisa denominada de aprendizagem de ontologias (*ontology learning)*, que é definida como um conjunto de métodos e técnicas para a construção semiautomática de uma nova ontologia ou para o enriquecimento de uma já existente.

Porém, para construir uma ontologia de maneira semiautomática é necessário a automatização do processo de aquisição de conhecimento e, para isso, diversas abordagens foram sugeridas (MAEDCHE; STAAB, 2002): construção de ontologias a partir de texto; construção de ontologias a partir de dicionários; construção de ontologias a partir de base de conhecimento; construção de ontologias a partir de esquemas semiestruturados; e construção de ontologias a partir de bases de dados relacionais.

Considerando a grande quantidade de documentos digitais disponíveis na *Web* e os que cada profissional e organizações possuem, o melhor seria considerá-los como um recurso válido para aquisição de conhecimento para a construção de ontologias. Dada a grande quantidade de textos disponíveis, espera-se que seja possível utilizar métodos que permitam construir uma ontologia que trate de uma parte razoável do domínio. Porém, a precisão e consistência não podem ser garantidas com métodos automatizados, sendo necessário sempre o pós-processamento por seres humanos.

Segundo Buitelaar et al. (2003) a utilização de textos como fonte de aquisição de conhecimento parece ser um caminho correto, visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos.

Face ao exposto, o presente projeto tem como questão norteadora: Pode uma ontologia na área da saúde ser construída de maneira semiautomática a partir de textos em português disponibilizados pelos profissionais ou disponíveis na *Web?*

1.1 OBJETIVO

1.1.1 Objetivo Geral

Desenvolver uma ferramenta para a construção semiautomática de ontologias a partir de textos em português.

1.1.2 Objetivos Específicos

- Implementar a ferramenta para construção semiautomática de ontologias.
- Validar os resultados extraídos pela ferramenta construída.

1.2 ESTRUTURA DO DOCUMENTO

Esta dissertação está estruturada nos seguintes capítulos: Introdução, Revisão de Literatura, Método, Resultados e Conclusões.

O primeiro capítulo é constituído pela introdução e objetivos. O segundo apresenta a revisão de literatura, abordando como temas centrais: ontologias, aprendizagem de ontologias e ferramentas para aprendizagem de ontologias. O terceiro capítulo refere-se às etapas do método.

Os resultados e discussão estão estruturados de acordo com as etapas do método, ou seja, desenvolvimento do protótipo; validação do protótipo; desenvolvimento da versão final da ferramenta; e validação dos termos extraídos pela ferramenta.

Finaliza-se com as conclusões e perspectivas de trabalhos futuros.

2 REVISÃO DE LITERATURA

Neste capítulo são apresentados, de forma breve, alguns conceitos importantes à compreensão deste estudo. As seções apresentam noções básicas da área de ontologias: definições, tipos, critérios para a construção de ontologias e exemplos de ontologias na área de saúde; aprendizagem de ontologias, com ênfase na aprendizagem a partir de textos; e as ferramentas existentes para aprendizagem de ontologias.

2.1 ONTOLOGIAS

Na área da Filosofia, Ontologia significa o estudo do ser ou da existência. Na área da Inteligência Artificial, os primeiros conceitos que surgiram definiram ontologias como termos e relações que compreendem o vocabulário de uma área, como também as regras para combinar estes termos e relações para definir extensões deste vocabulário.

Segundo Thomas Gruber "uma ontologia é uma especificação explícita de uma conceitualização" (GRUBER, 1993). Esta definição foi modificada, posteriormente, por Borst que complementou dizendo que "uma ontologia é definida como uma especificação formal de uma conceitualização compartilhada" (BORST, 1997).

Studer e seus colegas (1998) esclarecem esta definição:

- especificação explícita: podem-se entender as definições de conceitos, instâncias, relações, restrições e axiomas;
- formal: significa que é declarativamente definida, portanto, compreensível para agentes de software e sistemas computacionais;
- conceitualização: por tratar de um modelo abstrato de uma área de conhecimento ou de um universo limitado de discurso; e
- compartilhada: por tratar de um conhecimento consensual, seja uma terminologia comum da área modelada ou acordada entre os desenvolvedores dos agentes que se comunicam.

Além da definição de Gruber (1993), muitas definições e explicações de ontologias foram propostas na literatura.

Neches e colegas (1991) definem ontologia não apenas como um conjunto de condições e relacionamentos sobre um vocabulário, mas também regras, para combinar os termos e seus relacionamentos para definir extensões para o vocabulário.

Grüninger e Fox (1995) declaram que ontologia é uma descrição formal de entidades, propriedades, relacionamentos, regras e comportamentos.

Huhns e Singh (1997) afirmam que ontologia é uma representação do conhecimento de algum domínio, no qual está disponível para todos os outros componentes de um sistema de informação.

Uschold e Jasper (1999) afirmam que, apesar de uma ontologia ter várias formas, ela sempre inclui um vocabulário específico e alguns significados dos termos quando preciso. Incluem-se definições e relacionamentos de como os termos estão interligados, o que impõem uma estrutura de domínio coletiva, restringindo as possíveis interpretações dos termos.

Weiss (1999) define Ontologia como uma especificação dos objetos, conceitos e relacionamentos de uma dada área de interesse. Na sua definição, uma ontologia vai além de especificar uma taxonomia de classes (ou tipos), pois deve também descrever os relacionamentos.

Os componentes básicos de uma ontologia, segundo (GRUBER, 1993) são:

- classes: usadas com ampla abrangência, podendo expressar qualquer coisa sobre a qual alguma coisa é dita, como uma tarefa, função, ação, estratégia, processo de raciocínio, entre outros;
- relações: representam um tipo de interação entre as classes e o domínio. Como exemplo, pode-se considerar: “subclasse-de” ou “conectado-a”;
- funções: consistem de um caso especial de relações, onde o n-ésimo elemento do relacionamento é único para os n-1 elementos precedentes;
- axiomas: usados para modelagem de sentenças que são sempre verdadeiras, no universo modelado.

Ontologias podem ser de diferentes tipos, dependendo de fatores como o domínio pretendido a ser modelado, o uso para o qual são construídas ou a sua complexidade. A seguir são apresentadas algumas classificações.

2.1.1 Tipos de ontologias

Guarino (1998) classifica as ontologias de acordo com o seu nível de generalidade. Ele propõe a seguinte classificação:

- ontologias de alto nível: descrevem conceitos muito gerais ou conhecimento de senso comum, os quais são independentes de um problema ou domínio em particular;
- ontologias de domínio: descrevem o vocabulário relacionado a um domínio genérico;
- ontologias de tarefas: definem conceitos relacionados com a execução de uma determinada tarefa ou atividade;
- ontologias de aplicação: descrevem conceitos dependendo de um domínio específico ou uma tarefa em particular. Estas ontologias muitas vezes estendem e especializam ontologias de domínio e ontologias de tarefas.

Neste trabalho as estruturas ontológicas construídas de maneira semiautomática são consideradas de domínio.

Para Uschold e Grüninger (1996), ontologias são diferenciadas pelo grau de formalismo nos quais seus termos são escritos, classificando como:

- altamente informal: quando é expressa em linguagem natural. Este tipo de ontologia poderá ter muitas ambiguidades, devido à característica ambígua da linguagem natural;
- semi-informal: expressa em uma forma restrita e estruturada de linguagem natural, de forma a aumentar a clareza e diminuir a ambiguidade;
- semi-formal: expressa em linguagem artificial, formalmente definida;
- rigorosamente formal: os termos são meticulosamente definidos com semântica formal, teoremas e regras de propriedades.

Quanto ao grau de formalismo, as estruturas ontológicas construídas neste trabalho são consideradas semi-formais.

Independente da classificação da Ontologia, alguns princípios básicos para o seu desenvolvimento devem ser considerados (Gruber, 1993):

- clareza: a ontologia deve representar de maneira eficaz o significado dos termos definidos, os termos devem ser acompanhados de definições objetivas;
- coerência: a ontologia deve ser coerente, para permitir derivar inferências que sejam consistentes com as definições;
- extensibilidade: a ontologia deve ser construída levando em consideração a possibilidade de sua continuidade para permitir a inclusão de novos termos sem revisão das definições existentes;
- mínimo compromisso ontológico: para permitir que sejam definidas tão poucas suposições quanto possíveis sobre o mundo a ser modelado, permitindo que as especializações e instanciações da ontologia sejam definidas com liberdade;
- mínimo viés de codificação: a conceituação deve ser especificada no âmbito do conhecimento, ou seja, a escolha da representação não deve se basear somente na conveniência da notação ou implementação.

2.1.2 Ontologias na área de saúde

Ontologias podem ser modeladas em diversos domínios, dentre os quais o domínio da saúde. A seguir são apresentados alguns exemplos de ontologias na área de saúde.

Nome	Objetivo	Fonte
Unified Medical Language System	O objetivo é facilitar o desenvolvimento de sistemas computacionais que se comportem como se "entendessem" o significado da linguagem da biomedicina e saúde. Com este objetivo, base de dados e ferramentas de software com UMLS são produzidas e distribuídas para uso por desenvolvedores de sistemas, para construir ou melhorar sistemas de informação eletrônica que criem, processem, recuperem, integrem e/ou agreguem dados e informações biomédicos e da saúde.	(UMLS, 2009)

Nome	Objetivo	Fonte
Open Biomedical Ontologies	Experimento colaborativo envolvendo desenvolvedores de ontologias que estão estabelecendo princípios para o desenvolvimento de ontologias de referência no domínio biomédico.	(OBO, 2009)
Systemized Nomenclature of MEDicine	O SNOMED é uma ontologia médica que na versão inglesa contém cerca de 300.000 conceitos, 900.000 descrições e 1.300.000 relacionamentos entre conceitos, envolvendo quase todo o vocabulário usado na Medicina e naturalmente pelas aplicações médicas.	(SNOMED, 2009)
Gene Ontology	Fornece vocabulário estruturado e controlado sobre vários domínios da biologia molecular e celular. GO engloba os genes, produtos dos genes e suas sequências.	(GO, 2009)
ImMunoGeneTics	Primeira ontologia do domínio da imunogenética e foi criada para integrar bases de dados especialistas em antígenos receptores e o complexo maior de histocompatibilidade. Esta é a principal ontologia em uso pela comunidade científica do domínio.	(IMGT, 2009)
Major Histocompatibility Complex	Representa todos os genes do MHC humano e grupos sorológicos. Tem com objetivo facilitar a pesquisa de documentos na área e anotação de textos.	(MHC, 2009)
International Classification for Nursing Practice	O ICNP® foi desenvolvido em Web Ontology Language (OWL[1]) usando o editor de ontologias Protégé. O objetivo é facilitar o desenvolvimento do ICNP®, assim como a sua manutenção e implementação.	(ICNP, 2009)

Quadro 1: Ontologias na área de saúde
Fonte: Autor

Adquirir conhecimento para a construção de uma Ontologia requer muito tempo e recursos, pois é um trabalho minucioso onde o profissional deve pesquisar na literatura (ou consultar especialistas) termos relevantes, organizá-los hierarquicamente e identificar as relações existentes. Na tentativa de minimizar estas dificuldades surgiu a área de aprendizagem de ontologias.

2.2 APRENDIZAGEM DE ONTOLOGIAS

A Aprendizagem de Ontologias (*Ontology Learning)*, também conhecida como processo de Geração de Ontologias (*Ontology Generation*), ou Extração de Ontologias (*Ontology Extraction*), é definida como um conjunto de métodos e técnicas para a construção

[1] A OWL (Web Ontology Language) é uma linguagem para codificar ontologias na Web (SMITH; WELTY; MCGUINNESS, 2004).

de uma ontologia ou enriquecimento de uma já existente, utilizando-se de diversos tipos de fontes, como dicionários, textos ou base de dados.

Diversas abordagens foram sugeridas para a automatização do processo de aquisição de conhecimento, entre as quais se destacam a análise de linguagem natural e a aprendizagem de máquina. Segundo Maedche e Staab (2002) podem-se classificar as diversas abordagens da seguinte forma:

- construção de ontologias a partir de texto: ontologias são criadas através da aplicação das técnicas de análise de linguagem natural;
- construção de ontologias a partir de dicionários: ontologias são criadas utilizando processamento automático para ler os dicionários e extrair os termos relevantes e seus relacionamentos;
- construção de ontologias a partir de base de conhecimento: ontologias são criadas a partir de bases de conhecimento;
- construção de ontologias a partir de esquemas semiestruturados: ontologias são criadas a partir de fontes que têm alguma estrutura pré-definida, como por exemplo arquivos XML;
- construção de ontologias a partir de bases de dados relacionais: ontologias são criadas a partir da extração de conceitos e relações disponíveis em bases de dados.

Para cada um dos tipos de abordagem citados acima, existem técnicas específicas e algumas ferramentas relacionadas. Devido ao objetivo deste trabalho estar relacionado com a criação de ontologias a partir de textos, esta abordagem será detalhada a seguir.

2.2.1 Aprendizagem de Ontologias a partir de textos

Segundo Buitelaar e colegas (2005), a tarefa de construção de ontologias a partir de texto abrange basicamente as seguintes etapas: extração de termos, descoberta dos sinônimos, geração dos relacionamentos (não hierárquicos) e geração taxonômica.

A **extração dos termos** é um pré-requisito para todos os outros passos da aprendizagem de ontologias a partir de textos. Os termos estão intrinsecamente ligados ao domínio, e por isso esta etapa é a mais complexa de todas. Encontram-se diversos exemplos de métodos para extração de termos, alguns são baseados em recuperação de informação (*information retrieval*), em termos indexados, e outros baseados em terminologia e processamento de linguagem natural (PLN).

Na próxima fase, de **descoberta dos sinônimos** são recuperadas as variações semânticas dos termos. Muitos trabalhos sugeridos desta área usam o WordNet (WORDNET, 2008) para a aquisição dos sinônimos em inglês e o EuroWordNet (EUROWORDNET, 2008) para sinônimos de outras línguas do continente europeu.

Atualmente, existem diversos paradigmas para a **geração de taxonomias** a partir de textos, os três principais são: padrões baseados em informações léxico-sintáticas, baseada em termos complexos e a clusterização hierárquica dos termos. Estes três paradigmas são explicados a seguir.

Os padrões baseados em informações léxico-sintáticas propõem "uma abordagem que utiliza um conjunto de padrões baseados em informações léxico-sintáticas" (HEARST, 1992). O objetivo do método proposto é a extração de relações de hiponímia (relação "é-um") dos textos. Em Morin e Jacquemin (2004) também é proposto uma abordagem baseada em informações linguísticas, como ocorre em Hearst (1992), mudando apenas, em alguns casos, as estruturas extraídas e a língua alvo, no caso, a Francesa.

Na abordagem baseada em termos complexos, também chamados de compostos ou multi-palavras, o método analisa a estrutura interna dos termos, portanto, quando um termo está lexicamente contido em outro, ele é considerado seu hipônimo. Os trabalhos que utilizam esta técnica foram desenvolvidos por Buitelaar et al. (2003); Ryu; Choi (2006); Baségio (2008).

A clusterização hierárquica dos termos consiste na construção de estruturas taxonômicas a partir da clusterização conceitual e hierárquica dos termos. Faure e Nedellec (1999) propõem um método em que os clusters de base são formados por palavras que ocorrem com o mesmo verbo após a mesma preposição. Então é calculada a similaridade semântica entre os clusters, e os mais próximos são agrupados, formando novos clusters. Esse processo é repetido até restar somente um único cluster (raiz), onde todos os termos estão presentes.

Vários trabalhos na área de **geração dos relacionamentos (não hierárquicos)** sugerem uma combinação de mineração de texto (*text mining*), análises estatísticas e linguísticas, explorando as estruturas sintáticas para extrair relacionamentos entre os conceitos da ontologia (CIRAMITA et al., 2005).

A **extração de regras** é a etapa menos abordada na aprendizagem de ontologias. Alguns esboços de técnicas podem ser encontrados em (LIN; PANTEL, 2001).

O foco neste trabalho serão as etapas de extração dos termos e geração taxonômica para a construção semiautomática de ontologia. Nas próximas sessões estão descritas as abordagens encontradas na literatura para cada uma das duas fases.

2.2.2 Métodos para a extração de termos

Os principais métodos para extração de termos, os quais serão explicados a seguir são: Frequência relativa (FR), *Term frequency-inverse document frequency* (TF-IDF), *Log-likelihood e a Entropia*.

A **Frequência relativa (FR)** é uma das mais simples medidas encontrada na literatura. Ela considera o número de vezes em que um determinado termo aparece em um documento, dividido pelo total de termos do documento (MANNING; SCHÜTZE, 1999). Para calcular a medida Frequência relativa (FR) usa-se a fórmula (1):

$$FR(t) = \frac{F(t)}{N} \tag{1}$$

Onde,

- *F(t)* é o número de vezes em que o termo *t* aparece em um texto.
- '*N'* é o número total de termos no texto.

A ***Term frequency-inverse document frequency*** **(TF-IDF)** pesa a frequência de um termo em um documento. O termo que possuir uma alta frequência de ocorrência em um conjunto limitado de documentos é relevante para o domínio (MANNING; SCHÜTZE, 1999). Para calcular esta medida usa-se a fórmula (2):

$$tfidf_{l,d} = lef_{l,d} * \log(\frac{|D|}{dfl}) \quad (2)$$

Onde,

- $lef_{l,d}$ é a frequência da entrada lexical *'l'* em um documento *'d'*.
- *dfl* é o número de documentos do *corpus*[2] '*D*' em que *'l'* ocorre.
- '*D*' é a quantidade de documentos no *corpus*.

Para calcular o valor *tfidf* para uma entrada lexical no *corpus* inteiro, basta fazer o somatório das frequências desta entrada em cada documento do *corpus*. A fórmula (3) é utilizada para este cálculo:

$$tfidf_{l,d} = \sum_{d \in D} tfidf_{l,d} \quad (3)$$

Log-likelihood é uma medida que calcula a relevância de um termo do *corpus* do domínio com base na sua frequência no próprio *corpus* do domínio e na lista de referência (MANNING; SCHÜTZE, 1999). Para calcular a medida Log-Likelihood usa-se a fórmula (4):

$$G2 = 2 * ((a * \ln(\frac{a}{E1})) + (b * \ln(\frac{b}{E2}))) \quad (4)$$

Onde,

- $E1 = c * \frac{(a+b)}{(c+d)}$
- $E2 = d * \frac{(a+b)}{(c+d)}$
- *'a'* é a frequência do termo na lista de referência.
- '*b'* corresponde à frequência do termo no *corpus* de domínio.
- *'c'* indica o total de termos na lista de referência.
- *'d'* indica o total de termos no *corpus* de domínio.

Segundo Wiener (1948), "a soma de informação em um sistema é a medida de seu grau de organização; a **entropia** é a medida de seu grau de desorganização; um é o negativo

[2] *Corpus* é um conjunto amplo e estruturado de textos, os quais são usados para fazer uma análise estatística, verificação de ocorrências ou validação de regras linguísticas em um domínio específico.

do outro". Para calcular esta medida utilizam-se as fórmulas (5) e (6) adaptadas do Text-To-Onto (MAEDCHE; STAAB, 2003).

$$p_{l,d} = \frac{lef_{l,d}}{gf_{l,D}} \tag{5}$$

$$entropia_{l,d} = 1 - \frac{\sum_{d \in D} p_{l,d} * \log(p_{l,d})}{\log(D)} \tag{6}$$

Onde,

- $lef_{l,d}$ é a frequência da entrada lexical *'l'* em um documento *'d'*.
- $gf_{l,D}$ é o número de vezes em que a entrada lexical 'l' ocorre no *corpus* 'D'
- 'D' é a quantidade de documentos no *corpus*.

2.2.3 Métodos para geração taxonômica

Alguns métodos, com diferentes abordagens, foram propostos para geração taxonômica, dentre os quais estão os Padrões de Hearst, Padrões de Morin/Jacquemin e Abordagem Baseada em termos compostos.

Padrões de Hearst

O método **padrões de Hearst** (HEARST, 1992) utiliza um conjunto de padrões baseados em informações léxico-sintáticas. O objetivo do método proposto é a extração de relações de hiponímia (relação "é-um") dos textos. Este padrão foi criado para aplicação em textos da língua inglesa. O Quadro 2 apresenta os padrões propostos por Hearst.

Padrões
NP such as {(NP,)*(or|and)} NP
such NP as {(NP,)*(or|and)} NP
NP {, NP}* {,} or other NP
NP {, NP}* {,} and other NP
NP {,} including {NP,}*{or|and} NP

Quadro 2: Padrão de Hearst
Fonte: (HEARST, 1992)

Onde,

- NP = Noun Phrase - Sintagma Nominal
- | = separador de alternativas (ou)
- {} = agrupamento
- () = agrupamento
- * = zero ou mais ocorrências do elemento precedente

Abaixo segue uma lista de exemplos dos padrões léxico-sintáticos propostos por Hearst, aplicados em fragmentos de textos:

- NP such as {(NP,)*(or|and)} NP
  "... works by such authors as Herrick, Goldsmith, and Shakespeare."
  hiponímia ("author", "Herrick")
  hiponímia ("author", "Goldsmith")
  hiponímia ("author", "Shakespeare")

- NP {, NP}* {,} or other NP
  "Bruises, wounds, broken bones or other injuries ..."
  hiponímia ("bruise", "injury")
  hiponímia ("wound", "injury")
  hiponímia ("broken bone", "injury")

- NP {, NP}* {,} and other NP
  "...temples, treasuries, and other important civic buildings"
  hiponímia ("temple", "civic building")
  hiponímia ("treasury", "civic building")

- NP {,} including {NP,}*{or|and} NP
  "All common-law countries, including Canada and England ..."
  hiponímia ("Canada", "common-law country")
  hiponímia ("England", "common-law country")

- NP {,} especially {NP,}*{or|and} NP
  "...most European countries, especially France, England, and Spain"
  hiponímia ("France", "European country")
  hiponímia ("England", "European country")
  hiponímia ("Spain", "European country")

Padrões de Morin/Jacquemin

Outro método que utiliza padrões léxico-sintático foi proposto por Morin e Jacquemin (2004), porém, para a língua francesa. O Quadro 3 apresenta este padrão.

Padrões
{deux|trois...|2|3|4...} NP1 (LIST2)
{certain|quelque|de autre...} NP1 (LIST2)
{deux|trois...|2|3|4...} NP1: LIST2
{certain|quelque|de autre...} NP1: LIST2
{de autre}? NP1 tel que LIST2
NP1, particulièrement NP2
{de autre}? NP1 comme LIST2
NP1 tel LIST2
NP2 {et|ou} de autre NP1
NP1 et notamment NP2

Quadro 3: Padrão de Morin/Jacquemin
Fonte: (MORIN; JACQUEMIN, 2004)

Onde,

- NP = Noun Phrase (Sintagma Nominal)
- LIST = lista de sintagmas nominais separados por vírgula (",")
- | = separador de alternativas (ou)
- {} = agrupamento
- () = agrupamento
- ? = zero ou uma ocorrência do elemento precedente

Os padrões são aplicados da mesma forma que o padrão de Hearst (1992), mudando apenas em alguns casos as estruturas extraídas. Abaixo são apresentados alguns exemplos:

- {deux|trois...|2|3|4...} NP1 (LIST2)
  "... analyse foliaire de quatre espéces ligneuses (chêne, frêene, lierre et cornouiller) dans l'ensemble des sites étudiés."
  hiperônimo ("chêne", "espéces ligneuses")
  hiperônimo ("frêene", "espéces ligneuses")
  hiperônimo ("lierre", "espéces ligneuses")
  hiperônimo ("cornouiller", "espéces ligneuses")

- {certain|quelque|de autre...} NP1 (LIST2)
  "Aprés cinq années de résultats sur les principales cultures vivriéres (sorgho, maïs, mil), il apparaît qu'il existe un grand nombre de combinaisons possibles."
  hiperônimo ("sorgho", "cultures vivriéres")
  hiperônimo ("maïs", "cultures vivriéres")
  hiperônimo ("mil", "cultures vivriéres")

- {de autre}? NP1 tel que LIST2
  "Des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par des méthodes de routine."
  hiperônimo ("sodium", "cations")
  hiperônimo ("potassium", "cations")
  hiperônimo ("calcium", "cations")
  hiperônimo ("magnésium", "cations")

Abordagem baseada em termos compostos

Esta abordagem é baseada em termos complexos, também chamados de compostos ou multi-palavras. Esse método analisa a estrutura interna dos termos, portanto, quando um termo está lexicamente contido dentro de outro, ele é considerado seu hiperônimo. Alguns dos

trabalhos que utilizam esta técnica são Buitelaar et al. (2003); Ryu; Choi (2006); Baségio (2008). Na Figura 1 é ilustrado um exemplo para esta abordagem.

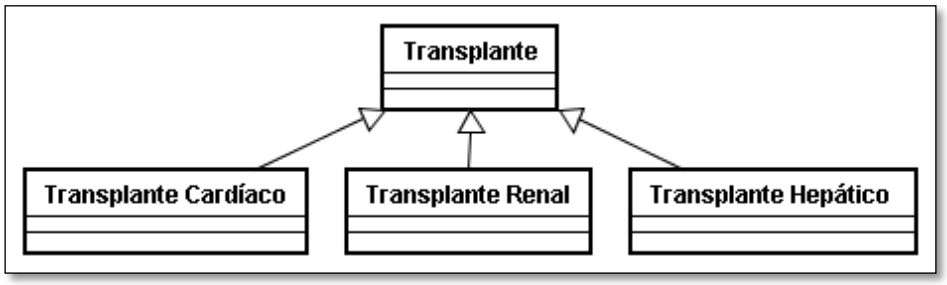


Figura 1: Exemplo da abordagem baseada em termos compostos
Fonte: Autor

Após o surgimento de diversas abordagens e técnicas para a aprendizagem de ontologias, ferramentas de aprendizagem de ontologias foram surgindo para a língua estrangeira e portuguesa.

2.3 FERRAMENTAS PARA CONSTRUÇÃO DE ONTOLOGIAS

2.3.1 Ferramentas para línguas estrangeiras

O Quadro 4, adaptado de Gómez-Pérez, Manzano-Macho (2003), apresenta um comparativo entre as ferramentas para aprendizagem de ontologias para línguas estrangeiras encontradas na literatura.

Nome	Objetivo e Escopo	Fonte de dados	Intervenção do usuário	Idioma do corpus	Fonte
ASIUM	Objetivo do ASIUM é encontrar relacionamentos taxonômicos entre textos em linguagem natural na língua francesa.	Texto analisado morfossintaticamente	Todo o processo	Francês	(FAURE; POIBEAU, 2000)
Caméléon	Objetivo do Caméléon é identificar padrões léxico-sintático ou utilizar o padrão de Hearst para encontrar relacionamentos não hierárquicos e relacionamentos taxonômicos.	Texto analisado morfossintaticamente	Validação ou definição de novas regras léxicas sintáticas	Inglês e Francês	(AUSSENAC-GILLES; SEGUELA, 2000)
CORPORUM-Ontobuilder	Objetivo do CORPORUM-Ontobuilder é encontrar relacionamentos taxonômicos entre textos em linguagem natural na língua inglesa.	Texto em linguagem natural	Não é necessário	Inglês	(ENGELS, 2001)

Nome	Objetivo e Escopo	Fonte de dados	Intervenção do usuário	Idioma do corpus	Fonte
DOE: Differential Ontology Editor	DOE é um editor de ontologias que permite o engenheiro criar ontologias em três passos de acordo com a metodologia proposta por Bruno Bachimont.	Texto em linguagem natural	Todo o processo	Inglês	(BACHIMONT, 1996)
KEA: Keyphrases Extraction Algorithm	KEA é uma ferramenta que extrai automaticamente palavras-chave dado um conjunto de textos utilizando técnicas como aprendizagem de máquina, processamento léxico e estatístico.	Texto em linguagem natural	Avaliação do resultado final	Inglês	(JONES; PAYNTER, 2002)
LTG (Language Technology Group) Text Processing Workbench	LTG é um conjunto de ferramentas que tem o objetivo de descobrir estruturas internas de textos na língua inglesa	Texto em linguagem natural	Todo o processo	Inglês	(MIKHEEV; FINCH, 1997)
Mo'K Workbench	Mo'K é uma ferramenta que tem como objetivo extrair conceitos taxonômicos utilizando anotações morfossintáticas, denominado pelo autor como "triplet".	Texto analisado morfossintaticamente	Todo o processo	Francês	(BISSON et al., 2000)
OntoLearn Tool	OntoLearn é uma ferramenta que tem como objetivo extrair termos de domínio de um texto, relacioná-los com conceitos adequados para uma ontologia e extrair relacionamentos entre eles. Para realizar estas tarefas ele utiliza técnicas como processamento de linguagem natural (PLN) e aprendizagem de máquina.	Texto em linguagem natural	Avaliação do resultado final	Inglês	(VELARDI et al., 2002)
Prométhée	Prométhée é uma ferramenta que tem como objetivo extrair e refinar padrões léxico-sintáticos relativos a relacionamentos conceituais, utilizando para realizar estas tarefas aprendizagem de máquina.	Texto em um padrão restrito	Todo o processo	Francês	(MORIN, 1999)
SOAT: a Semi-Automatic Domain Ontology Acquisition Tool	SOAT permite extrair de forma semiautomática ontologias de domínio usando como fonte *corpus* de domínio, identificando palavras-chave com ligações semânticas como sinônimos ou hiperônimos.	Texto em linguagem natural	Informações não disponíveis no artigo	Chinês	(WU et al.,2002)
SubWordNet Engineering Process tool	SubWordNet é uma ferramenta que tem como objetivo construir e manter línguas alternativas para o WordNet, utilizando técnicas como processamento de linguagem natural (PLN) e métodos estatísticos.	Texto em linguagem natural	Todo o processo	Inglês	(GUPTA et al., 2002)
SVETLAN'	SVETLAN' é uma ferramenta que cria clusters de palavras que aparecem dos textos fornecidos, independente de seu domínio. Ele utiliza técnicas como análise sintática e clusterização.	Texto em linguagem natural	Avaliação do resultado final	Francês	(CHAELANDA; GRAU, 2000)
TFIDF-based term classification system	TFIDF-based term classification system tem como objetivo extrair termos de domínio e relacionamentos entre eles, utilizando como técnica mineração de textos e métodos estatísticos.	Texto em linguagem natural	Avaliação do resultado final	Alemão	(XU et al.,2002)
TERMINAE	TERMINAE tem como objetivo auxiliar o engenheiro de ontologias na criação de uma ontologia, integrando ferramentas linguísticas e de engenharia do conhecimento.	Texto em linguagem natural	Avaliação do resultado final	Francês e Inglês	(BIÉBOW; SZULMAN, 1999)

Nome	Objetivo e Escopo	Fonte de dados	Intervenção do usuário	Idioma do corpus	Fonte
Text-To-Onto	Text-To-Onto tem como objetivo encontrar relacionamentos taxonômicos e não taxonômicos, utilizando técnicas como análise estatística, podas e regras de associação.	Texto em linguagem natural, dicionários e ontologias	Avaliação do resultado final	Inglês	(MAEDCHE; STAAB, 2003)
TextStorm and Clouds	O objetivo TextStorm and Clouds é a criação semiautomática de uma rede semântica, que contém apenas conceitos e seus relacionamentos, utilizando textos de um domínio específico.	Texto em linguagem natural	Todo o processo	Inglês	(NOVAK; GOWIN, 1984)
Welkin	Welkin é uma ferramenta que tem como objetivo gerar automaticamente material para um e-learning, utilizando técnicas como TFIDF e qui-quadrado.	Ontologia de domínio e WordNet	Não é necessário	Inglês	(ALFONSECA; RODRÍGUEZ, 2002)
WOLFIE (Word Learning From Interpreted Examples)	O objetivo do Wolfie é a aprendizagem léxico-semântica de um *corpus* de frases com representações de seus significados.	*Corpus* pré-processado com representações de seus significados	Avaliação do resultado final	Inglês	(THOMPSON; MOONEY, 1997)

Quadro 4: Comparativo entre ferramentas para aprendizagem de ontologias para línguas estrangeiras
Fonte: (GÓMEZ-PÉREZ, MANZANO-MACHO, 2003)(adaptado pelo Autor).

2.3.2 Ferramentas para língua portuguesa

Foram identificados na literatura dois trabalhos de aprendizagem de ontologias para língua portuguesa, os quais são propostos por Baségio (2008) e Ribeiro Junior (2008).

Baségio (2008) propõe uma abordagem para suportar as fases do processo de aquisição de estruturas ontológicas, mais especificamente as fases de extração de conceitos e relações taxonômicas, de modo à semi-automatizar os passos da construção de ontologias a partir de textos na língua portuguesa.

Para avaliação da abordagem proposta, foi desenvolvido um protótipo que incorpora mecanismos de importação de *corpus*, identificação de termos relevantes, identificação de relações taxonômicas entre esses termos e geração da estrutura ontológica em OWL. Este protótipo foi utilizado num estudo de caso sobre o domínio do Turismo. A Figura 2 ilustra uma visão simplificada desta abordagem.

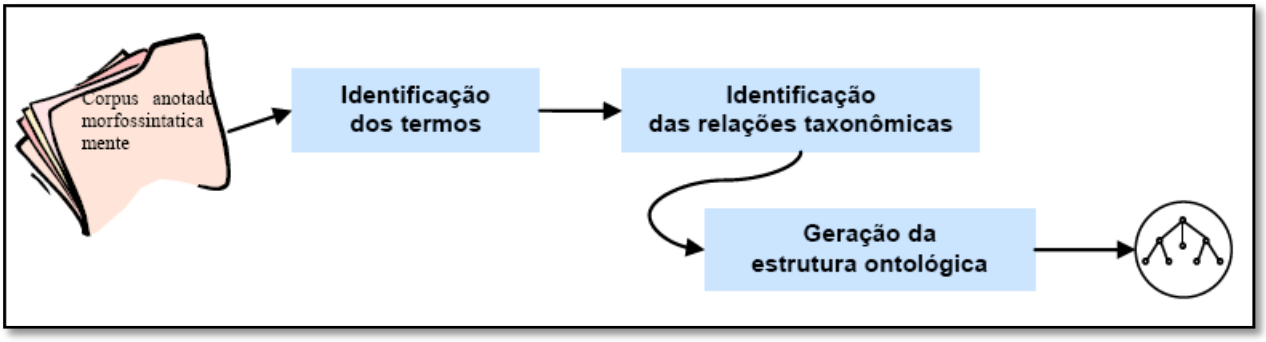


Figura 2: Abordagem proposta por Baségio.
Fonte: (BASÉGIO, 2008).

O método proposto utiliza como entrada um *corpus* anotado. De acordo com Baségio, as seguintes informações estão disponíveis nesse *corpus*:

- a palavra no seu formato original;
- o lexema da palavra original, ou seja, a palavra em sua forma singular e masculina; e
- a sua classe gramatical.

Para a identificação dos termos candidatos Baségio definiu as seguintes etapas:

- eliminação de termos que não representam conceitos de domínio: palavras que possuem significado semântico limitado são excluídas. Isto é feito através de uma lista de *stop words*[3];
- ponderação dos termos: nessa etapa são utilizadas as medidas *log-likelihood* e *tf-idf*;
- definição do limiar mínimo para os termos: o engenheiro de ontologia define um limiar mínimo para a medida *tf-idf*, termos que estiverem abaixo do ponto de corte são desconsiderados;
- exclusão/inclusão de termos: a lista resultante dos processos anteriores é apresentada ao engenheiro de ontologia, possibilitando que ele inclua ou exclua termos de acordo com seu julgamento;

[3] Lista padrão de palavras, símbolos ou caracteres irrelevantes ou proibidos.

- identificação de termos compostos: são selecionados termos compostos que tenham ao menos um termo considerado relevante em sua composição.

Para a tarefa de identificação de relações taxonômicas são utilizados os termos extraídos e avaliados anteriormente. Com base nesses termos são aplicados três métodos diferentes: identificação de relações taxonômicas com base em termos complexos, identificação de relações taxonômicas através dos padrões de Hearst (1992) e identificação de relações taxonômicas através dos padrões de Morin; Jacquemin (2004).

Em Ribeiro Junior (2008) é proposto um *plugin* para o ambiente de construção de ontologias Protégé. Este trabalho procura avançar nas questões referentes à tarefa para a língua portuguesa, abrangendo o desenvolvimento e a avaliação de métodos para a construção automática de ontologias a partir de textos.

O método proposto utiliza como entrada um *corpus* anotado no padrão XCES[4]. A Figura 3 exemplifica este padrão.

```
<?xml version="1.0" standalone="yes" ?>
- <cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
  - <struct type="token" from="0" to="1">
      <feat name="id" value="t1" />
      <feat name="base" value="O" />
    </struct>
  - <struct type="token" from="2" to="6">
      <feat name="id" value="t2" />
      <feat name="base" value="fumo" />
    </struct>
  - <struct type="token" from="7" to="8">
      <feat name="id" value="t3" />
      <feat name="base" value="é" />
    </struct>
  - <struct type="token" from="9" to="21">
      <feat name="id" value="t4" />
      <feat name="base" value="extremamente" />
    </struct>
  - <struct type="token" from="22" to="33">
      <feat name="id" value="t5" />
      <feat name="base" value="prejudicial" />
    </struct>
  - <struct type="token" from="34" to="35">
      <feat name="id" value="t6" />
      <feat name="base" value="a" />
    </struct>
  - <struct type="token" from="36" to="41">
      <feat name="id" value="t7" />
      <feat name="base" value="saúde" />
    </struct>
  - <struct type="token" from="41" to="42">
      <feat name="id" value="t8" />
      <feat name="base" value="." />
    </struct>
</cesAna>
```

Figura 3: Exemplo do formato XCES
Fonte: (RIBEIRO JUNIOR, 2008).

[4] XCES é um padrão baseado em XML para codificar *corpus*

Para a identificação dos termos candidatos Ribeiro Junior (2008) definiu as seguintes etapas:

- extração dos grupos semânticos; (opcional)
- filtragem dos grupos semânticos irrelevantes feita pelo engenheiro; (opcional)
- extração de termos simples considerando apenas aqueles pertencentes aos Grupos selecionados;
- exclusão dos termos simples irrelevantes realizada pelo engenheiro;
- extração dos termos complexos considerando apenas aqueles que possuem no mínimo uma palavra presente na lista final de termos simples e que pertençam a um grupo selecionado;
- exclusão dos termos complexos irrelevantes, feita pelo engenheiro.

Para o cálculo de relevância dos termos simples foram implementados os métodos: *FR, tf-idf e NC-Value* e para o cálculo de relevância dos termos complexos foram utilizadas os métodos *FR, tf-idf, C-Value e NC-Value.*

Para a tarefa de identificação de relações taxonômicas são utilizados os termos extraídos e avaliados anteriormente. Com base nesses termos são aplicados três métodos diferentes: identificação de relações taxonômicas com base em termos complexos, identificação de relações taxonômicas através dos padrões de Hearst e identificação de relações taxonômicas através dos padrões de Morin/Jacquemin.

3 MÉTODO

Esta pesquisa caracteriza-se como **aplicativa com abordagem quantitativa**. O método foi operacionalizado em **seis** etapas, apresentadas a seguir.

3.1 IDENTIFICAÇÃO DE FERRAMENTAS PARA APRENDIZAGEM DE ONTOLOGIA

Foi realizada uma busca na Web pelas ferramentas para aprendizagem de ontologia apresentadas no Quadro 3, com o objetivo de avaliá-las para identificar qual era a mais adequada para ser utilizada como base para este trabalho. Das dezoito ferramentas, apenas três estavam disponíveis para *download* e instalação, as quais foram instaladas e avaliadas. Utilizou-se, quando possível, um *corpus* com 15 artigos científicos sobre aprendizagem de ontologias, na língua inglesa.

Neste trabalho definiu-se avaliar as seguintes características: usabilidade (facilidade com que a interface pode ser utilizada), funcionalidade (resultados gerados de maneira satisfatória e corretamente) e portabilidade (facilidade de importação/exportação de ontologias em diversos formatos). Durante a avaliação, para cada característica foi atribuído um dos seguintes valores: 0 - não disponível; 1 - baixa presença; 2 - moderada presença; e 3 - alta presença.

Ao finalizar a avaliação das características as pontuações foram somadas sendo que este somatório correspondeu à nota atribuída para cada uma das ferramentas, onde à nota de maior valor corresponde a ferramenta com melhor avaliação.

Foram avaliadas as ferramentas KEA: Keyphrases Extraction Algorithm (JONES; PAYNTER, 2002), TERMINAE (BIÉBOW; SZULMAN, 1999) e Text-To-Onto (MAEDCHE; STAAB, 2003).

KEA: Keyphrases Extraction Algorithm

A ferramenta se propõe a fazer a extração automática de termos (chamada de palavras-chave pelo autor) para a construção de uma ontologia, utilizando um conjunto de

documentos com palavras-chave, selecionados manualmente por um especialista do domínio, como lista de referência e a medida *tfidf* como medida de extração.

Foi realizado o *download*, o desempacotamento e a importação do KEA no Eclipse. Sendo o projeto *open source* e sendo fornecidos os arquivos necessários para a importação no eclipse, o processo foi facilitado. Após esta etapa a ferramenta foi executada conforme o manual que é fornecido junto à ferramenta.

Além do código fonte, o KEA disponibiliza arquivos para realizar os testes da ferramenta. O Quadro 5 ilustra o resultado da execução da ferramenta com um dos arquivos teste fornecido, onde a primeira coluna apresenta as palavras-chave selecionadas automaticamente pela ferramenta e a segunda coluna as palavras-chave selecionadas manualmente por um especialista do domínio.

Extraídas Automaticamente	**Selecionadas Manualmente**
Bangladesh	**Bangladesh**
Cooperative farming	Development aid
Family planning	**Famine**
Famine	Food supply
First aid	Poverty
Food aid	Rural environment
Jute	Rural population
Landlessness	Social structure
Landowners	Tenure
Merchants	Villages

Quadro 5: Termos extraídos pelo KEA
Fonte: Autor

A ferramenta corresponde a sua proposta de extração de palavras-chave, mas a sua utilização é deficiente por dois motivos: a falta de uma interface gráfica e a necessidade de um especialista criar previamente uma lista de palavras-chave de um determinado domínio, dificultado o uso por usuários menos experientes. Tem-se como avaliação:

- usabilidade: “0 - não disponível”, por não existir uma interface gráfica;
- funcionalidade: “2 - moderada presença” pela necessidade de criar previamente uma lista de palavras-chave; e

- portabilidade: “0 - não disponível”, por não existir importação/exportação de ontologias em outros formatos.

TERMINAE

A ferramenta se propõe a fazer a extração semiautomática de termos, extraindo uma lista de termos candidatos e disponibilizando para o usuário selecionar os termos corretos.

Foi realizado o *download*, instalação e execução da ferramenta. A ferramenta está disponível apenas em Francês (Figura 4), dificultando o uso por pessoas que não conhecem esta língua. É também disponibilizado um manual em francês sobre a sua utilização. Para ser possível testar a ferramenta é preciso instalar uma ferramenta terceira utilizada para auxiliar o processo de extração de termos, denominada LEXTER. Entretanto, esta não está disponível para o *download*, limitando, assim, sua avaliação. Tem-se como avaliação:

- usabilidade: “1 - baixa presença”, pela interface gráfica ser confusa e pouco intuitiva;
- funcionalidade: “1 - baixa presença” pela necessidade de utilizar ferramentas terceiras não disponível para o *download*;
- portabilidade: “0 - não disponível”, por não existir importação/exportação de ontologias em outros formatos.

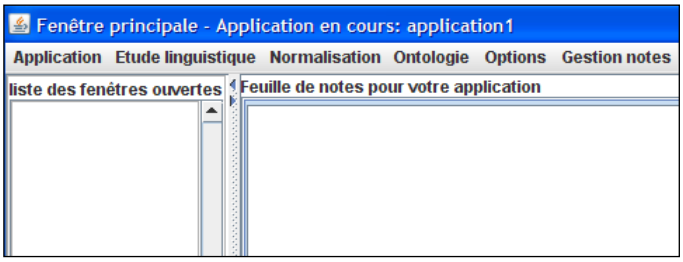


Figura 4: Interface gráfica do TERMINAE
Fonte: Autor

Text-To-Onto

A ferramenta se propõe a fazer a extração semiautomática de termos, extraindo uma lista de termos candidatos e disponibilizando para o usuário selecionar os termos corretos, a descoberta de relacionamentos (não hierárquicos) e a geração da taxonomia dos termos extraídos.

Foi realizado o *download* da ferramenta, o desempacotamento e a execução. É fácil de executar a ferramenta, pois estão disponíveis junto ao pacote de instalação, dois arquivos executáveis um de extensão sh (para ser executado no Sistema Operacional Linux) e um de extensão bat (para o Windows).

Neste experimento foi utilizado o *corpus* sobre aprendizagem de ontologias. A extração dos termos foi realizada utilizando quatro medidas: frequência, *tfidf*, entropia e c-value.

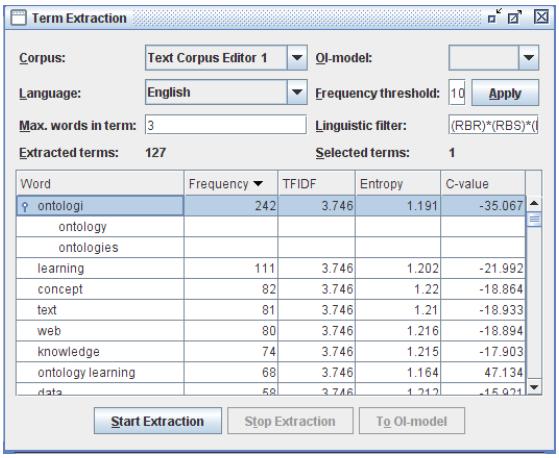


Figura 5: Termos extraídos no Text-to-Onto
Fonte: Autor

Com os termos extraídos (Figura 5) pode-se acionar outra funcionalidade chamada de *"Taxo Builder"* para criar automaticamente uma taxonomia com os termos processados anteriormente. Esta funcionalidade apresenta duas abordagens, a *Fca-based* e a *combination-based*, onde a ferramenta combina o padrão de Hearst com padrões heurísticos. A opção *Fca-based* não funcionou, travando o computador, sendo então utilizada a opção *combination-based*.

Outra funcionalidade analisada foi a de importar/exportar ontologias em outros formatos. A ferramenta disponibiliza apenas a opção de importar/exportar em formato RDF. Como experimento foi importada uma ontologia de exemplo disponibilizada pelo Protégé.

A ferramenta é intuitiva, sendo fácil aprender a utilizá-la, mas apresenta problemas de desempenho, travando o computador, mesmo com um *corpus* com apenas três documentos, que compromete sua funcionalidade. Tem-se como avaliação:

- usabilidade: "3 - alta presença", pela interface gráfica ser bem intuitiva e os menus estarem organizados de maneira a facilitar a assimilação;
- funcionalidade: "2 - moderada presença", pelos problemas de travamento que a ferramenta apresentou;
- portabilidade: "2 - moderada presença", por apenas disponibilizar a opção de importação/exportação em um formato, o RDF.

O Quadro 6 apresenta a avaliação final das ferramentas com a sua respectiva nota final. A ferramenta Text-to-Onto foi considerada a melhor para ser utilizada como base para criar uma ferramenta para construção semiautomática de ontologias em português. A KEA foi desconsiderada pela necessidade do especialista criar previamente uma lista de referência e a TERMINAE foi desconsiderada por não disponibilizar o *download* da ferramenta terceira utilizada para auxiliar o processo de extração de termos, dificultado sua avaliação.

Ferramenta	Usabilidade	Funcionalidade	Portabilidade	Nota
KEA	Não disponível	Moderada presença	Não disponível	2
TERMINAE	Baixa presença	Baixa presença	Não disponível	2
Text-To-Onto	Alta presença	Moderada presença	Moderada presença	7

Quadro 6: Avaliação final das ferramentas para aprendizagem de ontologia.
Fonte: Autor

3.2 IDENTIFICAÇÃO DE FERRAMENTAS DE ANOTAÇÃO LINGUÍSTICA

Esta etapa foi necessária devido à complexidade do desenvolvimento de uma ferramenta nova para anotação linguística, sendo necessária a avaliação das ferramentas gratuitas existentes para uso. Por se tratar de um projeto acadêmico, optou-se por utilizar apenas

ferramentas gratuitas. Foi realizado um levantamento das ferramentas de anotação linguística no portal do Linguateca (LINGUATECA, 2009), que é um portal que reúne ferramentas e publicações de processamento de linguagem natural (PLN) para a língua portuguesa. Selecionadas as ferramentas, foram realizados os respectivos *downloads* e testes para avaliar suas funcionalidades.

Foram encontradas três ferramentas e analisadas duas, pois uma delas está suspensa por tempo indeterminado. As características analisadas foram: facilidade de instalação, facilidade de uso (facilidade em aprender a utilizar a ferramenta) e interoperabilidade (possibilidade de utilizar a ferramenta de anotação linguística em conjunto com a ferramenta proposta); sendo atribuídos os valores: 0 - não disponível; 1 - baixa presença; 2 - moderada presença; e 3 - alta presença para cada uma das características avaliadas.

Ao finalizar a avaliação das características as pontuações foram somadas sendo que este somatório correspondeu à nota atribuída para cada uma das ferramentas, onde a nota de maior valor corresponde a ferramenta com melhor avaliação.

Foram avaliadas as ferramentas TreeTagger (SCHMID, 1994) e VISL (VISL, 2009).

TreeTagger

TreeTagger é uma ferramenta de anotação linguística, onde cada palavra é anotada com a sua classe gramatical e seu lema (forma canônica de uma palavra). Ela foi desenvolvida por Helmut Schmid (1994) no *Institute for Computational Linguistics of the University of Stuttgart* e suporta doze línguas diferentes, dentre elas, o português.

Após o *download* e o desempacotamento, foi executado o comando principal da ferramenta *"bin/tree-tagger.exe"*. Em seguida foi executada a linha: *tree-tagger.exe -token -lemma pt.par simulacao.txt simulacao_out.txt*, onde:

- o primeiro parâmetro é o arquivo executável;
- o segundo parâmetro "–token" é utilizado para incluir as palavras processadas no arquivo com o resultado final;
- o terceiro parâmetro "–lemma" é utilizado para incluir os lemas processados no arquivo com o resultado final;

- o quarto parâmetro é chamado pelo treetagger de *parameter file*, que é o *corpus* treinado na língua desejada, neste caso o português;
- o quinto parâmetro é o arquivo que será processado, o arquivo deve estar em um formato específico, onde as palavras devem estar cada uma em uma linha, como ilustrado na Figura 6;
- o sexto parâmetro corresponde ao arquivo onde o TreeTagger irá gravar o resultado final do processamento. A Figura 7 ilustra o resultado final.

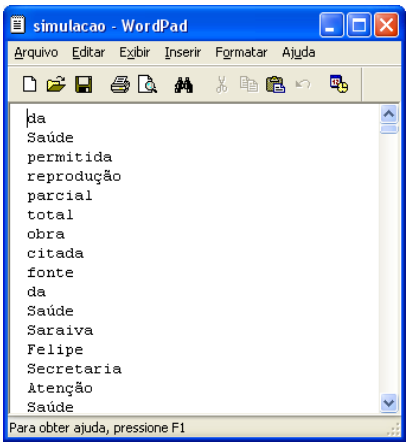


Figura 6: Arquivo de entrada para o TreeTagger
Fonte: Autor

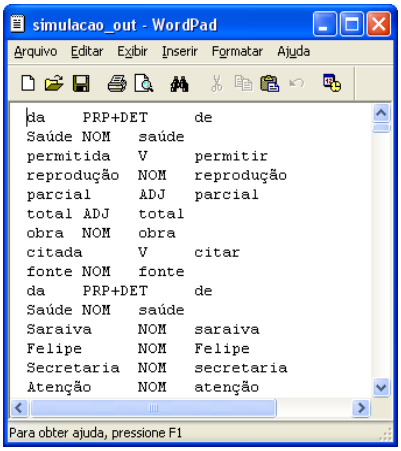


Figura 7: Arquivo de saída do TreeTagger
Fonte: Autor

Após realizar os testes tem-se a seguinte avaliação:

- facilidade de instalação: “3 - alta presença”, foi necessário apenas fazer um desempacotamento para instalar a ferramenta;

- facilidade de uso: “3 - alta presença”, pois a ferramenta dispõe de uma ajuda bem detalhada;
- interoperabilidade: “2 - moderada presença”, pois a ferramenta não disponibiliza nenhuma interface de comunicação, mas utilizando a ferramenta por linha de comando é possível utilizá-la de maneira integrada com outra ferramenta.

VISL - Visual Interactive Syntax Learning

O VISL é um projeto de pesquisa da *University of Southern Denmark* (SDU), desde 1996. Dentre as diversas funcionalidades ele disponibiliza um analisador morfossintático em quatorze línguas, dentre elas, o português.

O VISL disponibiliza três opções de uso: *online*, *upload* de arquivos e o serviço remoto cuja opção não é gratuita.

Primeiro testou-se a opção *online*, inserindo no campo de texto um artigo de apenas 16 páginas e 67 KB e a ferramenta apresentou um erro. Concluiu-se que este problema ocorreu devido ao tamanho do texto. Simulou-se então o uso da ferramenta utilizando apenas um parágrafo do mesmo artigo e a ferramenta se comportou de maneira normal cujo resultado é ilustrado na Figura 8.

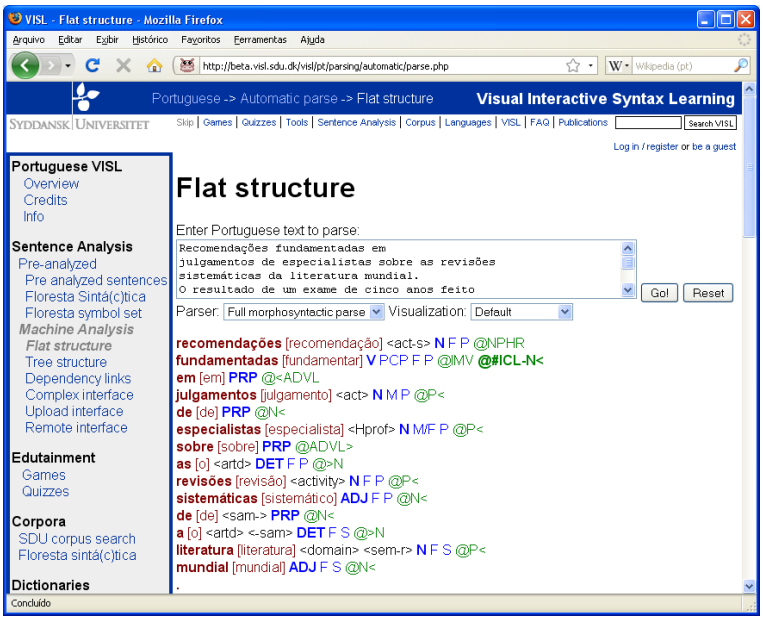


Figura 8: VISL em modo online
Fonte: Autor

O segundo passo foi testar a ferramenta em modo de *upload* de arquivos. Após cinco minutos um email foi enviado pela ferramenta com um *link* para o *download* do arquivo processado. Analisou-se então o resultado e foi possível concluir que a análise morfossintática foi realizada de maneira correta.

A terceira opção, serviço remoto, não foi possível testar, pois esta opção não é gratuita.

Depois de realizados os testes, a avaliação é:

- facilidade de instalação: "3 - alta presença", não foi necessário instalar, por se tratar de uma ferramenta *Web*.
- facilidade de uso: "3 - alta presença", apesar da ferramenta não disponibilizar uma ajuda, as descrições das opções são intuitivas.
- interoperabilidade: "1 - baixa presença", pois o serviço remoto não é gratuito, dificultando a integração com outra ferramenta.

Depois de analisadas as ferramentas e calculadas os suas notas (Quadro 7), a ferramenta selecionada foi o TreeTagger, pois o VISL não correspondeu com a necessidade de ser gratuita.

Ferramenta	Facilidade de instalação	Facilidade de uso	Interoperabilidade	Nota
TreeTagger	Alta presença	Alta presença	Moderada presença	8
VISL	Alta presença	Alta presença	Baixa presença	7

Quadro 7: Avaliação das ferramentas de anotação linguística.
Fonte: Autor

3.3 ELABORAÇÃO DO PROTÓTIPO

Após a seleção da ferramenta base e da ferramenta de anotação linguística, houve a necessidade do desenvolvimento de um protótipo, sendo definidos os seguintes requisitos:

- importar um *corpus*, de arquivos em formato texto, processado previamente no TreeTagger;

- inserir um filtro mínimo para selecionar a quantia mínima de vezes em que um termo se repete no texto para que ele seja extraído pelo protótipo;
- restringir a seleção dos termos por substantivos;
- processar o *corpus* e exibir ao usuário os termos simples e compostos, utilizando como medida de seleção frequência e *tfidf*; e
- selecionar os termos desejados e exportá-los para o formato XLS, para posterior manipulação.

O protótipo foi desenvolvido para plataforma *desktop* utilizando o padrão *swing* da linguagem Java, versão 6, através do IDE (do inglês ***Integrated Development Environment*** ou **Ambiente Integrado de Desenvolvimento**) Eclipse.

3.4 AVALIAÇÃO DOS RESULTADOS DO PROTÓTIPO

O protótipo foi avaliado utilizando-se como *corpus* 12 artigos científicos da área de Histocompatibilidade. Foram avaliadas as seguintes características: facilidade de uso do protótipo, facilidade de distribuição do protótipo e qualidade dos termos extraídos automaticamente.

Os testes foram realizados em um ambiente com plataforma Windows XP utilizando um computador Pentium 4 com 3.3GHz e 1GB de memória RAM.

3.5 ELABORAÇÃO DA VERSÃO FINAL DA FERRAMENTA

Após a identificação de novos requisitos na fase anterior, o protótipo foi adaptado para a plataforma *Web* com o desenvolvimento de uma nova versão da ferramenta, utilizando como base os requisitos do protótipo e os novos requisitos:

- automatização no processo de criação do *corpus*, com os seguintes requisitos:
  - a possibilidade de processar artigos em formato PDF;
  - a integração da ferramenta com a ferramenta de anotação linguística TreeTagger;

- facilidade de distribuição da ferramenta, optando-se por um sistema baseado na plataforma *Web*;
- inserção de um filtro mais refinado na seleção dos termos, optando pela adoção da lista de *Stop Words*;
- inserção de outra medida de seleção, a medida de Entropia;
- inserção de termos relacionados com os Descritores em Ciências da Saúde (DeCS), para facilitar a seleção dos termos pelo especialista;
- possibilidade de exportar os resultados para OWL para manipulação em um editor de ontologias;
- extração de sinônimos, para também facilitar a seleção dos termos pelo especialista; e
- geração taxonômica dos termos extraídos;

O DeCS é um vocabulário estruturado trilíngue criado pela Biblioteca Virtual em Saúde (BIREME) para uso na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais, assim como para ser usado na pesquisa e recuperação de assuntos da literatura científica em bases de dados (DeCS, 2009).

Para verificar se o termo extraído é um termo DeCS, foi feito o download do seu vocabulário de 2008 (DeCS, 2009) e inserida na ferramenta, para que esta possa comparar automaticamente a correspondência dos termos extraídos com a lista de DeCS.

A extração de sinônimos foi desenvolvida utilizando o OpenThesaurusPT (OPENTHESAURUSPT, 2009), que é um projeto *open source* para a construção de um dicionário de sinônimos para a língua portuguesa com 13.207 palavras e 4.072 grupos de sinônimos. Cabe ressaltar que esta lista de sinônimos não é especifica da área de saúde, podendo assim causar alguns problemas de interpretação. Foi realizada uma busca por um dicionário de sinônimos específico e gratuito sem sucesso, então optou-se em utilizar o OpenThesaurusPT, mesmo com as possíveis falhas por ser uma ferramenta não especifica de sinônimos na área de saúde.

A versão final da ferramenta foi desenvolvida em tecnologias de código aberto e gratuitos, dentre elas pode-se citar o novo padrão de desenvolvimento de interfaces *Web* adotada pela SUN chamada de Java Server Faces (JSF), sendo os fornecedores de framework

JSF o Apache MyFaces (MYFACES, 2009) e Jboss RichFaces (RICHFACES, 2009). Foi utilizada a versão 6 do *kit* de desenvolvimento JAVA e como IDE de desenvolvimento o Eclipse.

3.6 VALIDAÇÃO DOS TERMOS EXTRAÍDOS PELA FERRAMENTA

De acordo com (DELLSCHAFT; STAAB, 2007), existem duas maneiras para realizar a validação do resultado de uma ferramenta de aprendizagem de ontologias: avaliação manual por especialistas da área ou comparar o resultado obtido com uma ontologia de referência, denominada de padrão ouro (*gold standard*).

Optou-se pela validação manual por especialistas, devido à falta de um padrão ouro e a dificuldade de se criar um. A validação foi realizada com textos em português de três áreas da saúde: câncer de mama; eventos adversos pós-vacinação; e inventário vocabular de enfermagem, utilizando-se três critérios distintos, explicados nas próximas seções. Isto ocorreu devido a dificuldade em conseguir especialistas disponíveis para avaliar a extração dos termos. O ideal seria realizar vários experimentos com grupos diferentes de especialistas, em uma mesma área de domínio.

Os testes foram realizados em um ambiente com plataforma Windows XP utilizando um computador AMD Atlhon Dual-Core 2.71 GHz com 4GB de memória RAM.

3.6.1 Validação na área de câncer de mama

Nesta área foram processados 8 textos do Ministério da Saúde e 16 textos do Instituto Nacional do Câncer (INCA), totalizando 24 arquivos no *corpus*. Foi utilizado o valor 30 como filtro mínimo para termos simples e compostos, ou seja, o termo simples e composto deveriam aparecer no mínimo 30 vezes no *corpus* para que a ferramenta realizasse a extração de termos. Foi utilizado o filtro “Apenas Substantivos”, para que a ferramenta extraísse apenas substantivos.

Para validar se os termos extraídos correspondiam aos termos relevantes na área, foram consultados dois especialistas. Cada especialista recebeu uma planilha Excel (Figura 9)

com os termos extraídos e foi solicitado que eles assinalassem com um X os termos que eles consideravam relevantes. Um dos especialistas validou os termos e posteriormente apresentou a sua validação para o outro especialista, promovendo uma discussão dos termos discordantes. Os especialistas então definiram juntos quais os termos que deveriam ser considerados. Os especialistas escolhidos foram um médico ginecologista obstetra com oito anos de experiência e uma enfermeira com experiência prática de 21 anos, doutora em enfermagem em Saúde Pública e Coletiva e docente na área de Políticas de Saúde por 10 anos.

	A	B	C	D	E	F
1	**Palavra**	**Total**	**Tf-Idf**	**Entropy**	**Decs**	**Selecionado**
2	Termo A	56	10,75	1	FALSO	
3	Termo B	93	10,06	1,01	FALSO	
4	Termo C	31	8,67	1,17	VERDADEIRO	
5	Termo D	193	8,96	1,03	FALSO	
6	Termo E	68	8,26	1,13	FALSO	
7	Termo F	92	8,67	1,09	VERDADEIRO	
8	Termo G	32	10,06	1,02	FALSO	
9	Termo H	194	8,67	1,06	VERDADEIRO	
10						
11						

Figura 9: Exemplo de instrumento enviado ao especialista para validação
Fonte: Autor

3.6.2 Validação na área de eventos adversos pós-vacinação

Foi processado o Manual de Vigilância Epidemiológica de Eventos Adversos Pós-Vacinação (MINISTÉRIO DA SAÚDE, 2008), utilizando o valor 3 como filtro mínimo para termos simples e compostos, ou seja, o termo simples e composto deveria aparecer no mínimo 3 vezes no *corpus* para a ferramenta realizar a extração dos termos. Foi utilizado o filtro "Apenas Substantivos", solicitando para que a ferramenta extraísse apenas substantivos.

Foram consultados três especialistas para validar quais dos termos extraídos eram relevantes na área. Da mesma forma que no experimento anterior, cada especialista recebeu uma planilha Excel (Figura 9) com os termos extraídos e foi solicitado que eles assinalassem com um X os termos que eles consideravam relevantes.

Foram considerados validados os termos que obtiveram uma porcentagem de 66% de concordância, ou seja, onde dois dos três especialistas concordaram com o termo, e os demais foram desconsiderados.

Os especialistas escolhidos foram dois profissionais da área de saúde que trabalham há mais de cinco anos com análise de eventos adversos pós-vacinação e um especialista há mais de dez anos em imunização.

O manual possui um glossário de termos e como estes termos são considerados relevantes e deveriam fazer parte da ontologia, não houve necessidade de serem validados pelos especialistas. Sendo assim, os termos do glossário foram comparados automaticamente com os termos extraídos pela ferramenta para avaliar a correspondência e retirados do processo de validação.

3.6.3 Validação na área de inventário vocabular de enfermagem

Um texto referente ao cenário da força de trabalho de Agentes de Saúde foi processado, cujo objetivo original com este texto era a contribuição para uma norma mundial de enfermagem em saúde coletiva (GARCIA; NÓBREGA, 2000). Foi utilizado o valor 2 como filtro mínimo para termos simples e 1 como filtro mínimo para termos compostos, ou seja, o termo simples deveria aparecer no mínimo 2 vezes no *corpus*, e o termo composto 1 vez, para que a ferramenta realizasse as suas extrações. Foi utilizado o filtro "Apenas Substantivos", solicitando para que a ferramenta extraísse apenas substantivos.

Para validar se os termos extraídos correspondiam aos termos relevantes na área, foi consultado um especialista que recebeu uma planilha Excel (Figura 9) com os termos extraídos e foi solicitado que ele assinalasse com um X os termos que ele considerasse relevantes.

O especialista selecionado foi uma enfermeira, bacharel em letras, com trinta e seis anos de experiência na área de sistematização da assistência de enfermagem, processo de enfermagem, sistemas de classificação, nomenclaturas, vocabulários, terminologias de enfermagem e representante do Centro de Pesquisa e Desenvolvimento da CIPE® Brasil.

4 RESULTADOS E DISCUSSÃO

Os resultados estão apresentados em quatro partes, sendo: protótipo, validação do protótipo, versão final da ferramenta e validação dos resultados extraídos pela ferramenta.

4.1 PROTÓTIPO

Um protótipo da ferramenta para construção semiautomática de ontologias em português foi elaborado. A Figura 10 ilustra a tela inicial do protótipo.

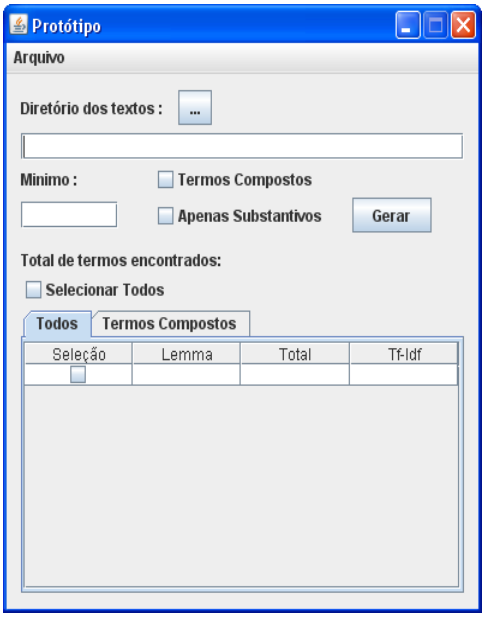


Figura 10: Protótipo
Fonte: Autor

O processo de geração semiautomática de ontologias do protótipo foi dividido em duas etapas: a **criação do *corpus*** e a **criação da ontologia**.

A **criação do *corpus*** (Figura 11) não foi implementada no protótipo, sendo nesta fase manual, onde o usuário deve processar cada documento de forma unitária e manual no TreeTagger. Caso os documentos estejam em formato PDF, o usuário deve transformá-los, previamente, em arquivo texto.

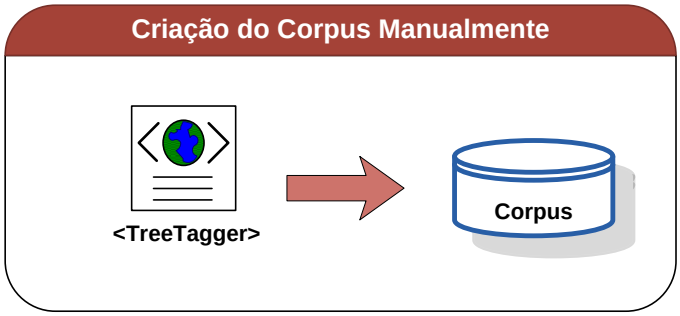


Figura 11: Processo de criação do *corpus* do Protótipo
Fonte: Autor

Com o *corpus* construído manualmente, a etapa de **criação da ontologia** pode ser iniciada, a qual foi dividida em cinco fases, descritas a seguir, conforme a Figura 12:

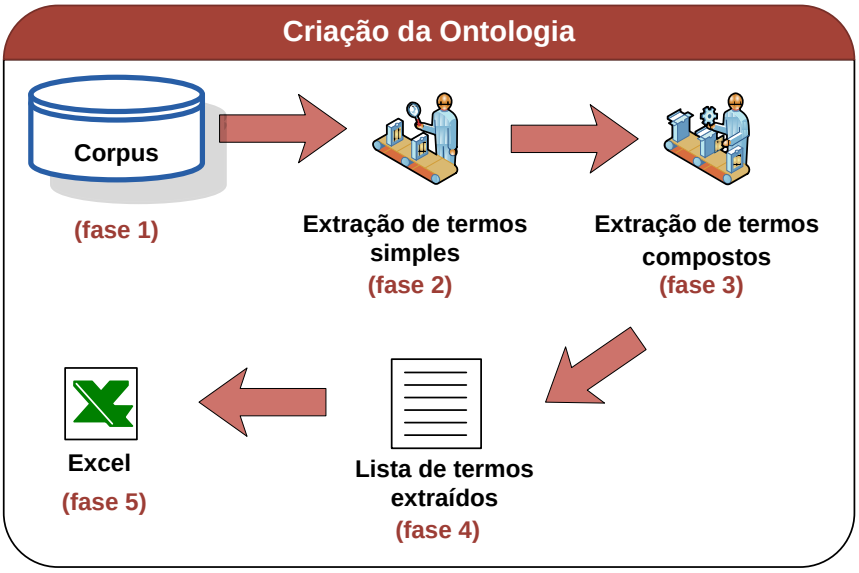


Figura 12: Processo da criação da ontologia do Protótipo
Fonte: Autor

1. O usuário seleciona o diretório que contém os artigos processados manualmente na etapa de **criação do *corpus*** (Figura 13) e preenche os filtros:
    - mínimo: quantidade mínima de vezes em que um termo simples, termo formado por apenas uma palavra, aparece no *corpus*;
    - termos compostos: incluir ou não no resultado os termos compostos;
    - apenas substantivos: incluir no resultado apenas termos marcados pelo TreeTagger como substantivos ou todos os termos.

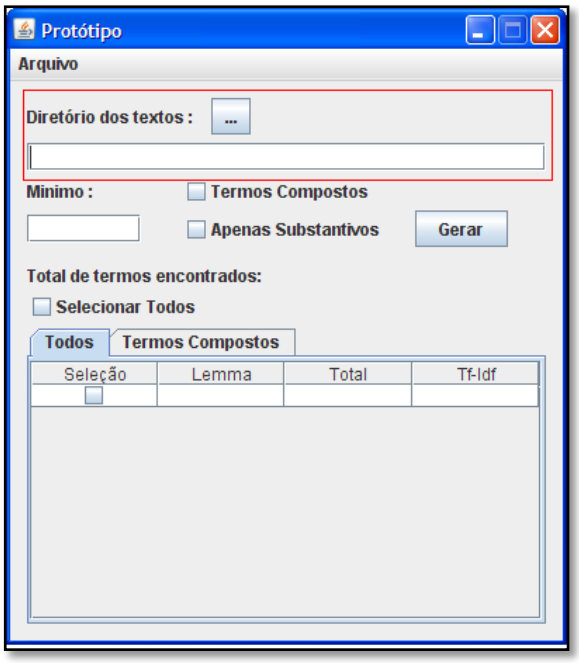


Figura 13: Importar *Corpus*
Fonte: Autor

2. Os termos simples são extraídos, aplicando as métricas frequência e *tfidf*, tendo como resultado a interface apresentada na Figura 14, onde a primeira coluna deve ser selecionada caso o usuário considere este termo relevante para a ontologia; a segunda coluna refere-se ao lema, que é a forma canônica da palavra; e a terceira coluna apresenta a métrica *tfidf*.

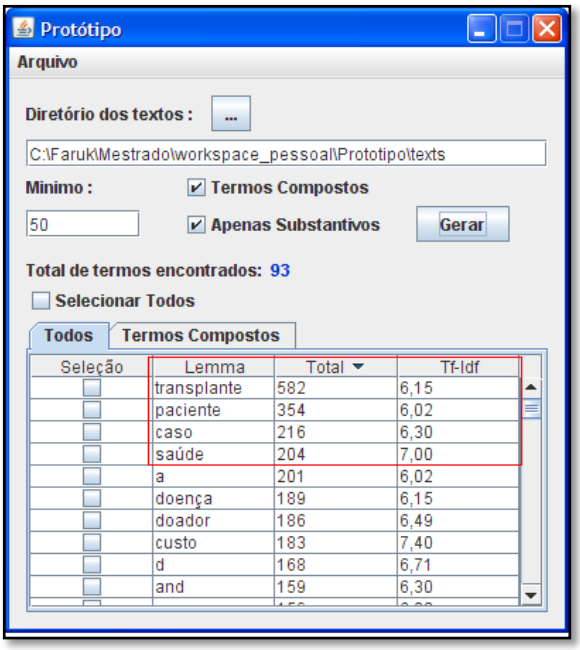


Figura 14: Extração de termos simples no Protótipo
Fonte: Autor

3. Os termos compostos são extraídos com base em regras expressas por sequência de tipos morfológicos de acordo com o Quadro 8, sendo aplicadas as medidas frequência e *tfidf*, tendo como resultado a interface apresentada na Figura 15, onde o significado das colunas é o mesmo descrito no item 3.

Regras de sequência morfológica
Substantivo + Adjetivo

Quadro 8: Regras de identificação de termos compostos no Protótipo
Fonte: Autor

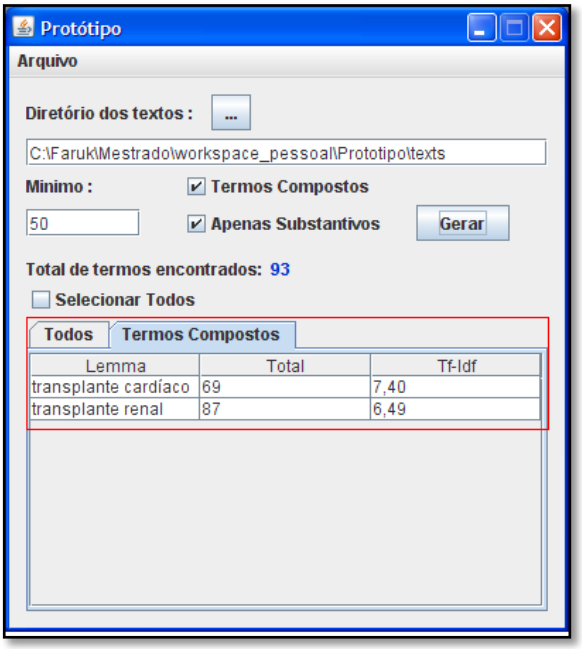


Figura 15: Extração de termos compostos no Protótipo
Fonte: Autor

4. Os termos simples e os compostos, extraídos pela ferramenta, são apresentados para o usuário, para que seja feita a seleção dos termos mais relevantes. O usuário deve selecionar na primeira coluna os termos que farão parte da ontologia (Figura 16).

5. O usuário pode exportar o resultado para o formato XLS, selecionando os termos relevantes ou a opção "selecionar todos", e em seguida a opção exportar para XLS (Figura 16). Para esta tarefa utilizou-se o *framework*

Apache POI (APACHE POI, 2009), que é uma biblioteca para trabalhar com arquivos dos formatos do *Microsoft Office* (MICROSOFT OFFICE, 2009).

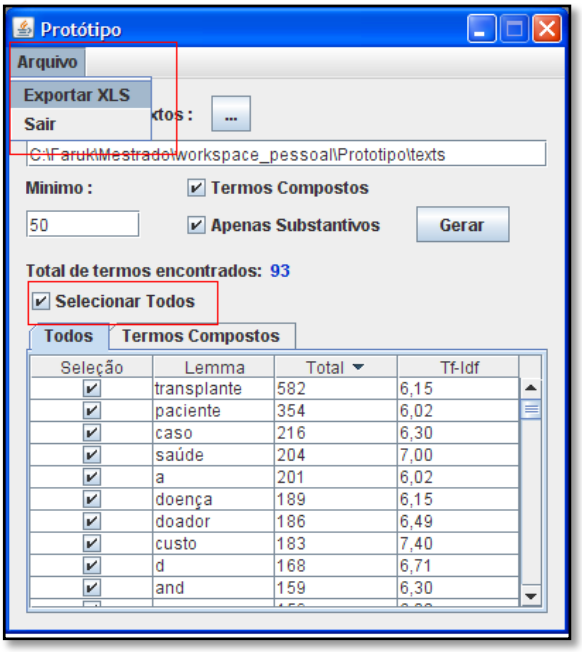


Figura 16: Exportar para XLS
Fonte: Autor

4.2 VALIDAÇÃO DO PROTÓTIPO

Para validação do protótipo desenvolvido algumas simulações foram realizadas, utilizando como *corpus* artigos científicos da área de Histocompatibilidade, com o objetivo de avaliar as seguintes características: facilidade de uso do protótipo, facilidade de distribuição e qualidade dos termos extraídos automaticamente.

Após testar o protótipo algumas melhorias foram identificadas, dentre elas:

- necessidade de envio dos arquivos em formato PDF facilitando o uso do usuário que não precisaria transformá-lo em arquivo texto manualmente;
- integração com a ferramenta de anotação linguística TreeTagger para a geração do *corpus* de entrada, pois no protótipo este passo era feito manualmente, fora da ferramenta;
- dificuldade de distribuição de uma versão *desktop*, optando por uma nova versão *Web*;

- necessidade de exportar os resultados para OWL para manipulação em um editor de ontologias;
- necessidade de um filtro mais refinado dos termos, optando pela exclusão de termos na lista de *Stop Words*;
- inserção da outra medida de seleção selecionada, a medida de entropia;
- inserção de termos relacionados com os Descritores em Ciências da Saúde (DeCS), para facilitar a seleção dos termos pelo especialista;
- seleção de sinônimos, para também facilitar a seleção dos termos pelo especialista; e
- geração da taxonomia.

4.3 VERSÃO FINAL DA FERRAMENTA

Após a identificação das melhorias no protótipo, foi possível desenvolver a versão final da ferramenta. A Figura 17 apresenta a página principal da ferramenta.

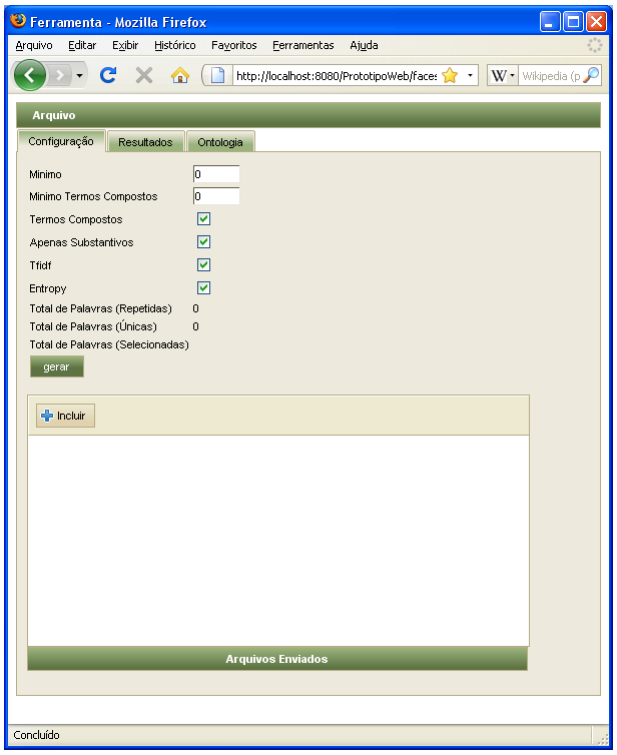


Figura 17: Página de entrada
Fonte: Autor

O processo de geração semiautomática de ontologias foi dividido em duas etapas: a **criação do *corpus*** e a **criação da ontologia**.

A **criação do *corpus*** (ilustrada pela Figura 18) é dividida em cinco fases:

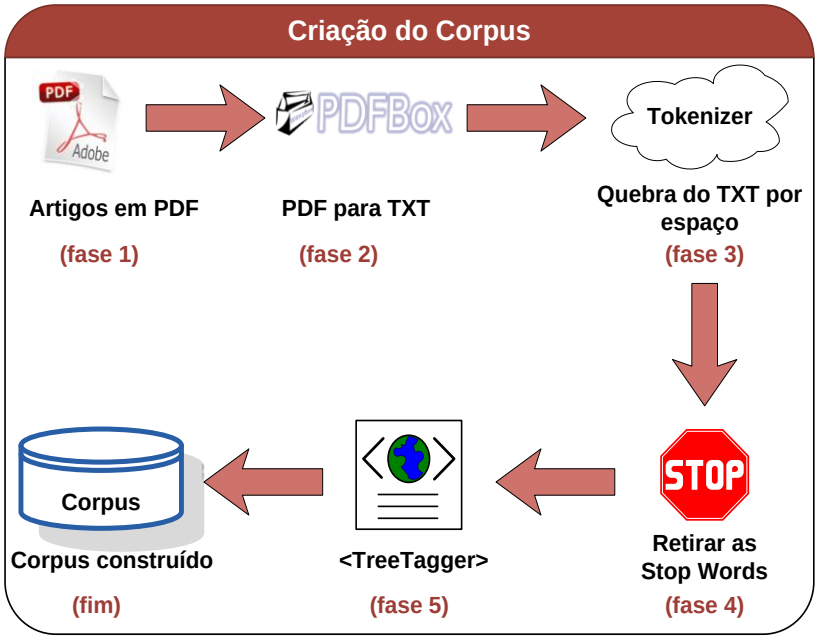


Figura 18: Processo de criação do *corpus* da Ferramenta
Fonte: Autor

1. O usuário envia os artigos em formato PDF que deseja processar (Figura 19)

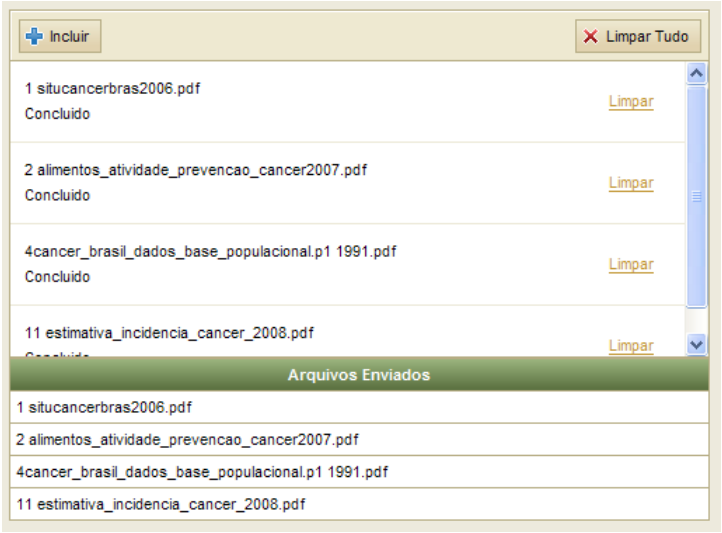


Figura 19: Envio dos artigos para processamento
Fonte: Autor

2. Os artigos são transformados em arquivos texto limpo (sem os marcadores padrões de arquivos PDF), utilizando o *PDFBox* (PDFBOX, 2009) que é uma biblioteca *open source* distribuída pela *Apache Incubator* para manipulação de documentos PDF, que permite criar novos arquivos PDF, manipulá-los e extrair o seu conteúdo.
3. Ocorre um pré-processamento dos textos, onde o texto é dividido por espaços em branco para posteriormente ser feito o processamento de anotação linguística com o TreeTagger.
4. As *Stop Words* são removidas.
5. Os textos são processados com o TreeTagger.

Com o *corpus* construído, a etapa de **criação da ontologia** (ilustrada pela Figura 20) pode ser iniciada, a qual é dividida em oito fases:

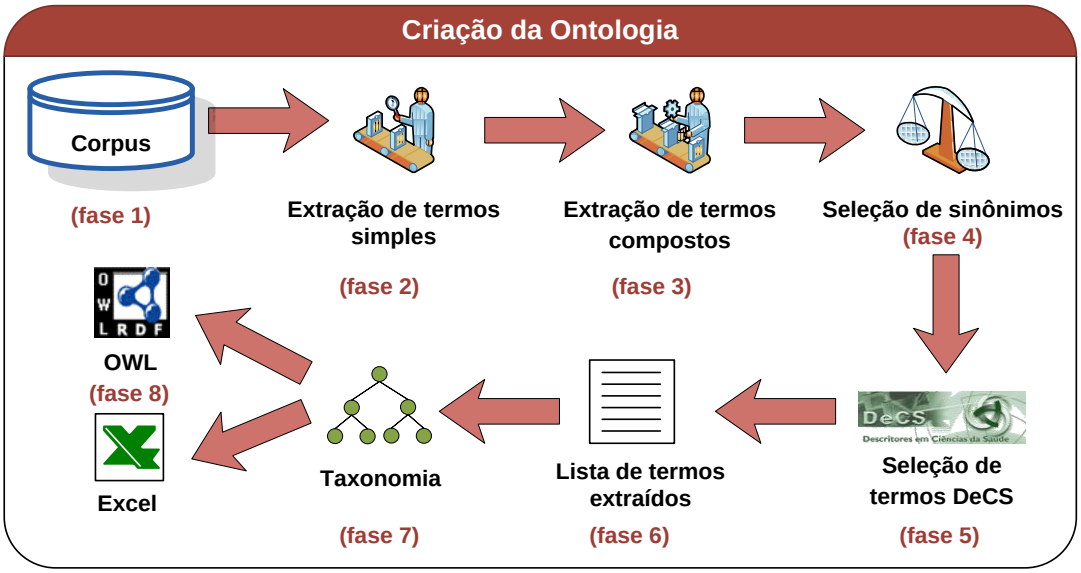


Figura 20: Processo da criação da ontologia da Ferramenta
Fonte: Autor

1. O usuário preenche os filtros:
    - mínimo: quantidade mínima de vezes em que um termo simples aparece no *corpus;*
    - mínimo de termos compostos: quantidade mínima de vezes em que um termo composto aparece no *corpus;*

- termos compostos: incluir ou não no resultado os termos compostos;
- apenas substantivos: incluir no resultado apenas termos marcados pelo TreeTagger como substantivos ou todos os termos;
- Tfidf: utilizar ou não a medida *tfidf* como medida de seleção;
- Entropy: utilizar ou não a medida de entropia como medida de seleção.

2. Os termos simples são extraídos aplicando as medidas frequência, *tfidf* e entropia.
3. Os termos compostos são extraídos com base em regras expressas por sequência de tipos morfológicos de acordo com o Quadro 9 e são aplicadas as medidas frequência, *tfidf* e entropia.

Regras de sequência morfológica
Substantivo + Adjetivo
Substantivo + Preposição + Substantivo
Substantivo + Preposição + Adjetivo + Substantivo
Substantivo + Preposição + Substantivo + Preposição + Substantivo

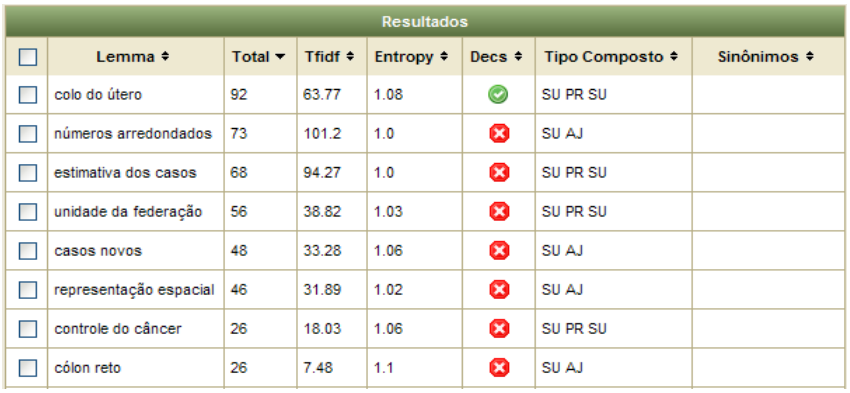
Quadro 9: Regras de identificação de termos compostos na Ferramenta
Fonte: Autor

4. Uma busca por sinônimos dos termos é realizada na lista do OpenThesaurusPT (OPENTHESAURUSPT, 2009), esta busca é realizada para facilitar o critério de seleção do termo pelo usuário .
5. É realizada uma pesquisa para verificar se os termos extraídos possuem correspondência na lista de Descritores em Ciência da Saúde (DeCS), novamente para facilitar o critério de seleção do termo pelo usuário.
6. Os termos simples (Figura 21) e os compostos (Figura 22), extraídos pela ferramenta, assim como alguns sinônimos para estes termos (Figura 23) e se o termo está ou não incluído na lista de descritores, são apresentados para o usuário, para que seja feita a seleção dos termos mais relevantes (Figura 24) que devem ser inseridos na ontologia.

Resultados

Lemma	Total	Tfidf	Entropy	Decs	Sinônimos
câncer	746	214.62	1.11	✓	câncer...
caso	485	139.53	1.04	✗	acontecimento...
taxa	386	111.05	1.05	✗	contribuição...
bruto	313	216.96	1.01	✗	atroz...
estimativa	291	83.72	1.03	✗	
localização	285	81.99	1.03	✓	
incidência	270	77.68	1.07	✓	

Figura 21: Extração de termos simples na Ferramenta
Fonte: Autor



Resultados

Lemma	Total	Tfidf	Entropy	Decs	Tipo Composto	Sinônimos
colo do útero	92	63.77	1.08	✓	SU PR SU	
números arredondados	73	101.2	1.0	✗	SU AJ	
estimativa dos casos	68	94.27	1.0	✗	SU PR SU	
unidade da federação	56	38.82	1.03	✗	SU PR SU	
casos novos	48	33.28	1.06	✗	SU AJ	
representação espacial	46	31.89	1.02	✗	SU AJ	
controle do câncer	26	18.03	1.06	✗	SU PR SU	
cólon reto	26	7.48	1.1	✗	SU AJ	

Figura 22: Extração de termos compostos na Ferramenta
Fonte: Autor

Resultados

Lemma	Total	Tfidf	Entropy	Decs	Sinônimos
câncer	746	214.62	1.11	✓	câncer...
caso	485	139.53	1.04	✗	a
taxa	386	111.05	1.05	✗	c
bruto	313	216.96	1.01	✗	atroz...

- câncer
- cancro

Figura 23: Sinônimos
Fonte: Autor

Resultados

Lemma	Total	Tfidf	Entropy	Decs	Sinônimos
câncer	746	214.62	1.11	✓	câncer...
caso	485	139.53	1.04	✗	acontecimento...
taxa	386	111.05	1.05	✗	contribuição...
bruto	313	216.96	1.01	✗	atroz...
estimativa	291	83.72	1.03	✗	
localização	285	81.99	1.03	✓	
incidência	270	77.68	1.07	✓	
novo	223	154.58	1.02	✗	actual...
saúde	220	63.3	1.08	✓	
brasil	217	62.43	1.12	✓	
neoplasia	215	61.86	1.05	✗	
pele	205	58.98	1.06	✓	couro...
primária	200	138.63	1.01	✗	
mulher	177	50.92	1.08	✗	dama...
tabela	168	116.45	1.03	✗	tabela...
melanoma	146	202.4	1.0	✓	
homem	141	40.57	1.07	✗	homem...
pulmão	130	37.4	1.08	✓	
risco	123	85.26	1.07	✓	linha...
população	105	30.21	1.11	✓	

« « 1 2 3 4 5 6 7 8 9 10 » »»

Figura 24: Lista dos termos para seleção
Fonte: Autor

7. Após selecionar os termos relevantes, o usuário pode optar por solicitar a organização destes termos em uma taxonomia (Figura 25 e Figura 26), que é construída com um método baseado em termos compostos, abordado na seção 2.2.3.

Figura 25: Menu Gerar ontologia
Fonte: Autor

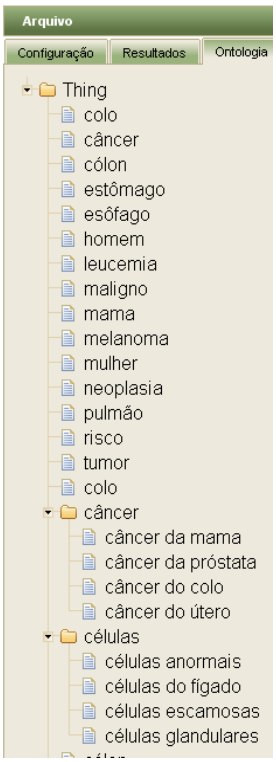
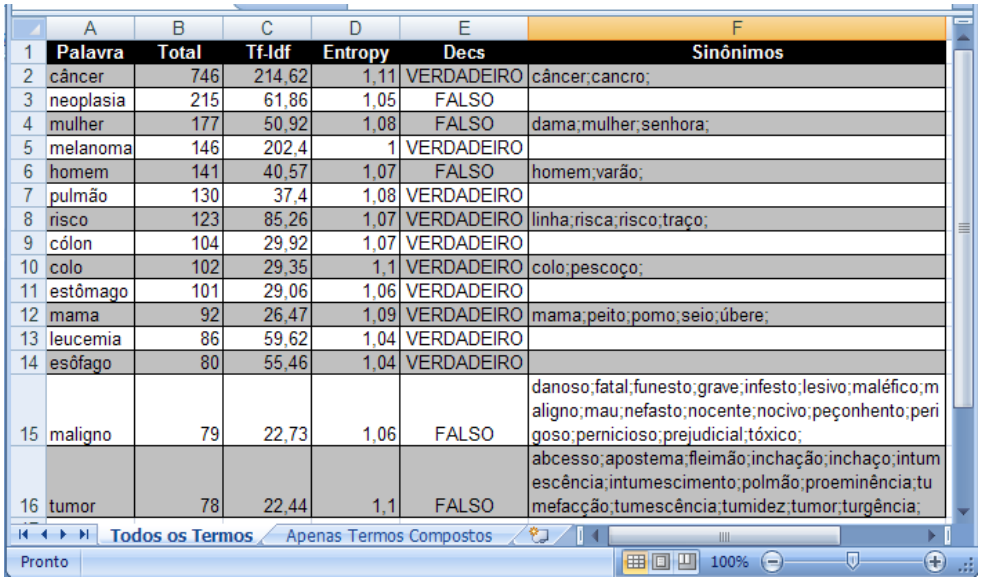


Figura 26: Taxonomia gerada
Fonte: Autor

8. O usuário pode exportar o resultado para o formato XLS ou OWL, selecionando a opção do Menu “Exportar para...”. A Figura 27 apresenta um exemplo de arquivo XLS gerado, sendo que para esta tarefa utilizou-se o *framework* Apache POI. A Figura 28 apresenta um exemplo de arquivo OWL gerado, o qual poderá ser manipulado em um editor de ontologias, como por exemplo, o Protégé (Figura 29). Para esta tarefa utilizou-se o *framework* Jena, que é uma biblioteca para manipulação de arquivos no padrão OWL, dentre outros.



	A	B	C	D	E	F
1	**Palavra**	**Total**	**Tf-Idf**	**Entropy**	**Decs**	**Sinônimos**
2	câncer	746	214,62	1,11	VERDADEIRO	câncer;cancro;
3	neoplasia	215	61,86	1,05	FALSO	
4	mulher	177	50,92	1,08	FALSO	dama;mulher;senhora;
5	melanoma	146	202,4	1	VERDADEIRO	
6	homem	141	40,57	1,07	FALSO	homem;varão;
7	pulmão	130	37,4	1,08	VERDADEIRO	
8	risco	123	85,26	1,07	VERDADEIRO	linha;risca;risco;traço;
9	cólon	104	29,92	1,07	VERDADEIRO	
10	colo	102	29,35	1,1	VERDADEIRO	colo;pescoço;
11	estômago	101	29,06	1,06	VERDADEIRO	
12	mama	92	26,47	1,09	VERDADEIRO	mama;peito;pomo;seio;úbere;
13	leucemia	86	59,62	1,04	VERDADEIRO	
14	esôfago	80	55,46	1,04	VERDADEIRO	
15	maligno	79	22,73	1,06	FALSO	danoso;fatal;funesto;grave;infesto;lesivo;maléfico;maligno;mau;nefasto;nocente;nocivo;peçonhento;perigoso;pernicioso;prejudicial;tóxico;
16	tumor	78	22,44	1,1	FALSO	abcesso;apostema;fleimão;inchação;inchaço;intumescência;intumescimento;polmão;proeminência;tumefacção;tumescência;tumidez;tumor;turgência;

Todos os Termos | Apenas Termos Compostos

Pronto 100%

Figura 27: Arquivo XLS gerado
Fonte: Autor

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <owl:Class rdf:ID="colo"/>
  <owl:Class rdf:ID="câncer"/>
  <owl:Class rdf:ID="células"/>
  <owl:Class rdf:about="#células escamosas">
    <rdfs:subClassOf rdf:resource="#células"/>
  </owl:Class>
  <owl:Class rdf:ID="melanoma"/>
  <owl:Class rdf:ID="risco"/>
  <owl:Class rdf:ID="esôfago"/>
  <owl:Class rdf:ID="pulmão"/>
  <owl:Class rdf:ID="estômago"/>
  <owl:Class rdf:about="#câncer da mama">
    <rdfs:subClassOf rdf:resource="#câncer"/>
  </owl:Class>
  <owl:Class rdf:ID="mulher"/>
  <owl:Class rdf:ID="mama"/>
  <owl:Class rdf:about="#câncer da próstata">
    <rdfs:subClassOf rdf:resource="#câncer"/>
  </owl:Class>
  <owl:Class rdf:ID="leucemia"/>
  <owl:Class rdf:about="#câncer do colo">
    <rdfs:subClassOf rdf:resource="#câncer"/>
```

Figura 28: Arquivo OWL gerado
Fonte: Autor

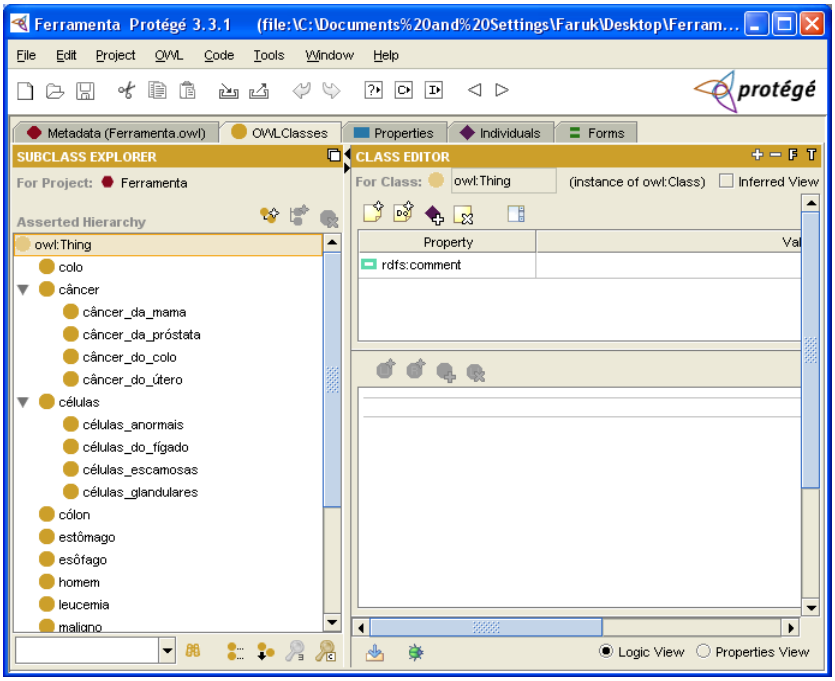


Figura 29: Ontologia importada para o Protégé
Fonte: Autor

4.4 VALIDAÇÃO DOS TERMOS EXTRAÍDOS PELA FERRAMENTA

Foram realizados experimentos com textos em português das áreas de câncer de mama (experimento 1); eventos adversos pós-vacinação (experimento 2); e inventário vocabular de enfermagem (experimento 3).

A execução dos experimentos iniciou com o processamento do *corpus* pela ferramenta. Foi contabilizada a quantidade de termos repetidos, ou seja, termos que se repetiram mais de uma vez no *corpus*; a quantidade de termos únicos, ou seja, sem levar em consideração a quantidade de vezes em que aparecem no *corpus*; e a quantidade dos termos do *corpus* extraídos pela ferramenta. O Quadro 10 apresenta estes valores.

	Câncer de mama	**Eventos Adversos pós-vacinação**	**Inventário vocabular de enfermagem**
Total de termos repetido no *corpus*	481.683	35.465	3.309
Total de termos únicos no *corpus*	55.207	6.169	681
Total de termos extraídos pela ferramenta	1.442 (2,61%)	1.090 (17,67%)	428 (62,85%)

Quadro 10: Síntese dos termos repetidos, termos únicos e termos extraídos
Fonte: Autor

Estes resultados se devem à poda realizada pelos filtros que eliminaram termos simples e compostos com frequência menor que o parâmetro inserido no filtro e termos simples que não eram substantivos. O Quadro 11 apresenta os parâmetros dos filtros.

	Câncer de mama	**Eventos Adversos pós-vacinação**	**Inventário vocabular de enfermagem**
Filtro mínimo de termos simples	30	3	2
Filtro mínimo de termos compostos	30	3	1
Apenas Substantivos	Sim	Sim	Sim

Quadro 11: Parâmetros para os filtros utilizados
Fonte: Autor

Os termos extraídos pela ferramenta foram enviados para os especialistas, que fizeram a seleção dos termos. Com isso foi possível contabilizar a quantidade de termos extraídos corretamente pela ferramenta. Foram selecionados 449 termos (APÊNDICE I) no experimento 1; 370 termos (APÊNDICE II) no experimento 2; e 190 termos (APÊNDICE III) no experimento 3, ou seja, a ferramenta extraiu corretamente 31,14%; 33,94%; e 44,39% em cada experimento respectivamente (Gráfico 1).

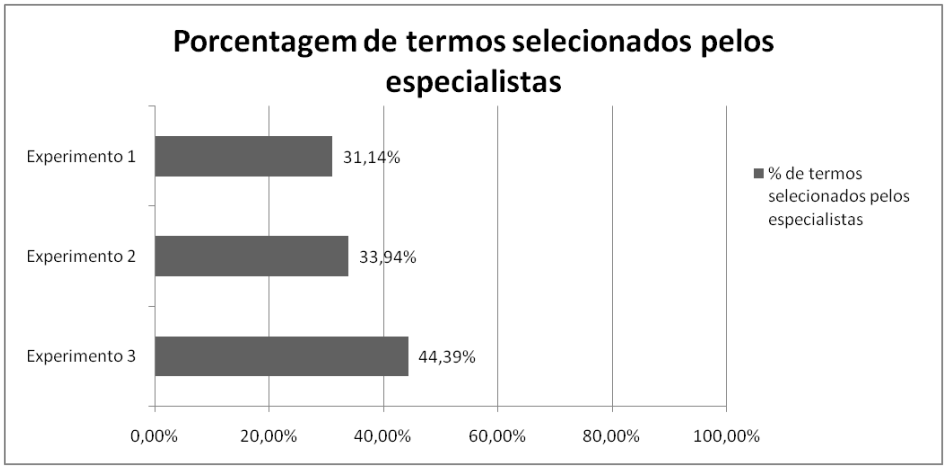


Gráfico 1: Comparação das porcentagens de termos extraídos corretamente
Fonte: Autor

No experimento 2, dos 1.090 termos candidatos, 620 foram selecionados pelo especialista 1, 161 pelo especialista 2 e 590 pelo especialista 3 (Gráfico 2). Após a seleção pelos três especialistas calcularam-se quais termos obtiveram uma porcentagem de 66% de concordância, ou seja, onde dois dos três especialistas concordaram com o termo, tendo como resultado 370 termos selecionados.

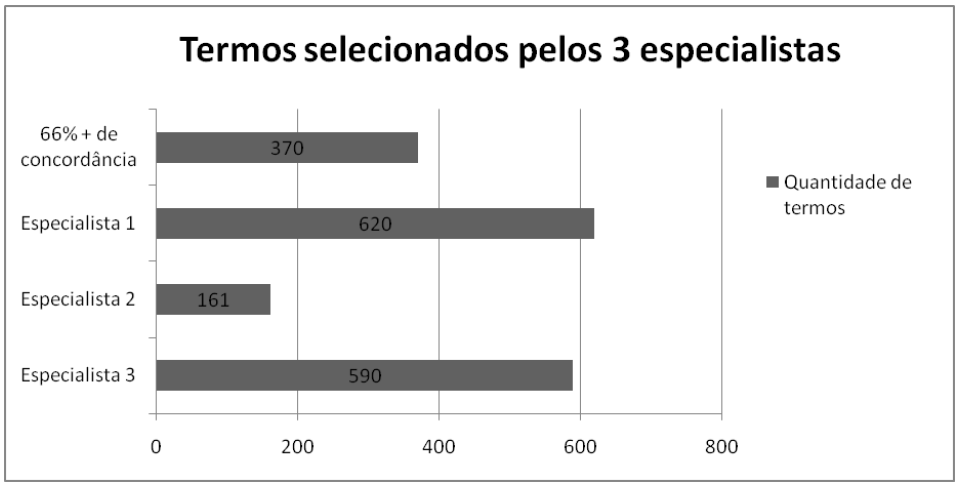


Gráfico 2: Total de termos selecionados pelos três especialistas do experimento 2
Fonte: Autor

O Quadro 12 apresenta alguns exemplos de termos que não foram selecionados por todos os especialistas. Pode-se notar que o especialista 1 selecionou mais termos que os outros dois especialistas, mas para que o termo fosse considerado correto, era preciso que pelo menos dois especialistas o selecionassem.

Termo	Quantidade	Especialista 1	Especialista 2	Especialista 3
pósvacinação	202	X		
manual	182	X		
secretaria	179	X		
hora	130	X		
reacção	126			X
manifestação	117	X		
anexo	83	X		
quadro	80	X		
célula	32	X		
hemograma	8			X

Quadro 12: Termos não selecionados pelos especialsitas no experimento 2
Fonte: Autor

A diferença entre a quantidade de termos selecionados pelos especialistas destaca a subjetividade do processo. É possível que o especialista 1 conheça mais sobre o domínio, tendo assim mais confiança na seleção dos termos, ou o especialista 2 e o especialista 3 serem

mais rigorosos quanto ao critério de seleção. Não é escopo deste trabalho analisar os critérios de seleção de termos, mas salientar a discussão sobre a subjetividade envolvida dentro do mesmo domínio e escopo.

No experimento 1 foram selecionados pelos especialistas apenas 31,14% dos termos extraídos devido à quantidade de "ruídos" no *corpus*, que são termos que não têm relevância para o domínio em questão. O Quadro 13 apresenta alguns exemplos de termos que não foram selecionados por nenhum dos especialistas.

Termo	Quantidade
brasil	1620
bruto	1546
tabela	1335
pele	1265
rio	1256
controle	1234
janeiro	1087
instituto	1068
leucemia	1030
rio de janeiro	1015

Quadro 13: Termos não selecionados pelos especialistas no experimento 1
Fonte: Autor

Pode-se observar que termos como "Brasil" e "Rio de Janeiro" foram extraídos e não foram selecionados. Isto ocorre porque a ferramenta não diferencia nomes próprios de substantivos. Isso indica a necessidade da criação de um filtro que possibilite a exclusão de nomes próprios do *corpus*.

Cabe destacar que o termo "leucemia", que para uma pessoa leiga parece ser um termo correto a ser selecionado, não foi selecionado pelos especialistas da área. Este fato salienta a importância da participação do especialista na etapa de pós-processamento e a dificuldade em se construir ontologias automaticamente. A análise do especialista é subjetiva e de acordo com um domínio e escopo específico, capacidade esta que, atualmente, a máquina não possui.

Com isso concluiu-se, como destacado em (DELLSCHAFT; STAAB, 2007), que os fatores que mais influenciam a avaliação são: a escolha correta dos especialistas e a quantidade deles, pois escolhendo com critério os especialistas tem-se a certeza de que os termos serão selecionados corretamente e com uma quantidade maior de especialistas eliminam-se as possíveis falhas geradas por eles.

Após a seleção dos termos pelos especialistas foi possível contabilizar a quantidade destes que são DeCS (Quadro 14). Nos três experimentos mais de 20% dos termos selecionados são DeCS. Com isso concluiu-se nos três experimentos que os descritores têm um peso fundamental para a extração dos termos automaticamente pela ferramenta, podendo ser utilizada como medida de extração.

	Câncer de mama	**Eventos Adversos pós-vacinação**	**Inventário vocabular de enfermagem**
Total de termos selecionados pelos especialistas	449 (31,14%)	370 (33,94%)	190 (44,39%)
Total de termos selecionados pelos especialistas que são DeCS	110 (24,50%)	144 (38,92%)	49 (25,79%)

Quadro 14: Quantidade dos termos selecionados que são DeCS
Fonte: Autor

O Quadro 15 apresenta a distribuição dos termos selecionados, nos três experimentos, levando em consideração a medida de frequência.

Pode-se concluir que mais da metade dos termos selecionados pelos especialistas aparecem com menor frequência no corpus (Gráfico 3). Isso sugere a criação de um limiar de corte com o máximo de vezes em que um termo pode aparecer em um *corpus*, além do mínimo de vezes utilizado neste trabalho.

	Câncer de mama		**Eventos adversos pós-vacinação**		**Inventário vocabular de enfermagem**	
Frequência dos termos selecionados	30 - 100 vezes	289 (64,37%)	3 - 50 vezes	342 (92,43%)	1 - 5 vezes	160 (84,21%)
	100 - 1000 vezes	148 (32,96%)	50 - 100 vezes	13 (3,51%)	5 - 10 vezes	23 (12,11%)
	mais de 1000 vezes	12 (2,67%)	mais de 100 vezes	15 (4,05%)	mais de 10 vezes	7 (3,68%)

Quadro 15: Distribuição dos termos selecionados levando em consideração a medida de frequência
Fonte: Autor

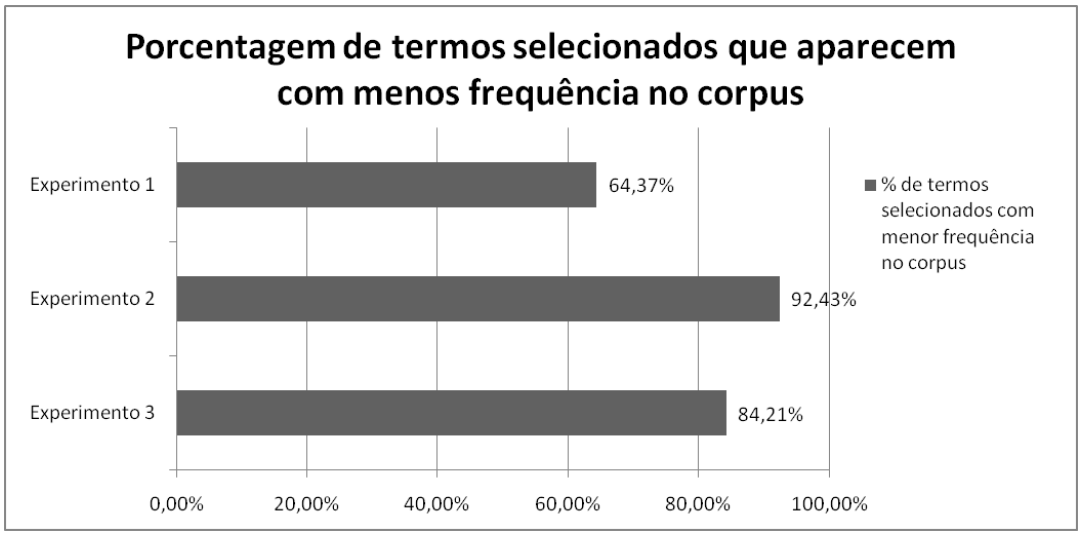


Gráfico 3: Porcentagem de termos selecionados pelos especialistas que aparecem com menor frequência no corpus
Fonte: Autor

O Gráfico 4 apresenta a quantidade de termos compostos extraídos e selecionados em cada experimento.

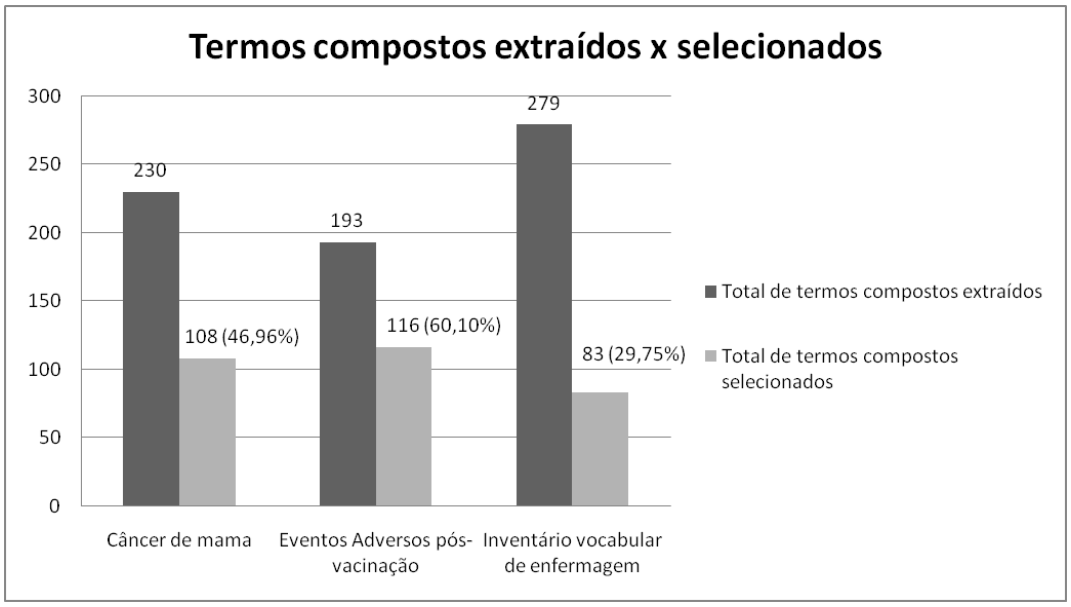


Gráfico 4: Termos compostos extraídos e selecionados
Fonte: Autor

Após a contabilização dos termos compostos extraídos e selecionados, pode-se contabilizar também a quantidade de termos compostos extraídos e selecionados de acordo com cada regra. O Quadro 16 apresenta a porcentagem de acerto no experimento 1, o Quadro 17 no experimento 2, o Quadro 18 no experimento 3.

Câncer de mama			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AJ	95	48	50,53%
SU PR SU	135	60	44,44%
SU PR AD SU	0	0	0,00
Total	**230**	**108**	**-**

Quadro 16: Porcentagem de acerto em cada regra de extração de termos compostos no experimento 1
Fonte: Autor

Ferramenta			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AJ	153	106	69,28%
SU PR SU	40	10	25,00%
SU PR AD SU	0	0	0,00
Total	**193**	**116**	**-**

Quadro 17: Porcentagem de acerto em cada regra de extração de termos compostos no experimento 2
Fonte: Autor

Ferramenta			
Regra	**Extraídos**	**Selecionados**	**Total de acertos**
SU AJ	92	28	30,43%
SU PR SU	147	46	31,29%
SU PR AD SU	0	0	0,00
SU PR SU PR SU	40	9	22,50%
Total	**279**	**83**	**-**

Quadro 18: Porcentagem de acerto em cada regra de extração de termos compostos no experimento 3
Fonte: Autor

O Gráfico 5 ilustra uma comparação dos termos compostos selecionados por cada regra nos três experimentos. As regras "SU AJ" e “SU PR SU" obtiveram um acerto maior que as regras "SU PR AD SU" e “SU PR SU PR SU[5]”.

[5] Esta regra foi desenvolvida por sugestão do especialista do experimento 3, e está presente apenas neste experimento, pois foi implementada na ferramenta após a execução dos outros dois experimentos.

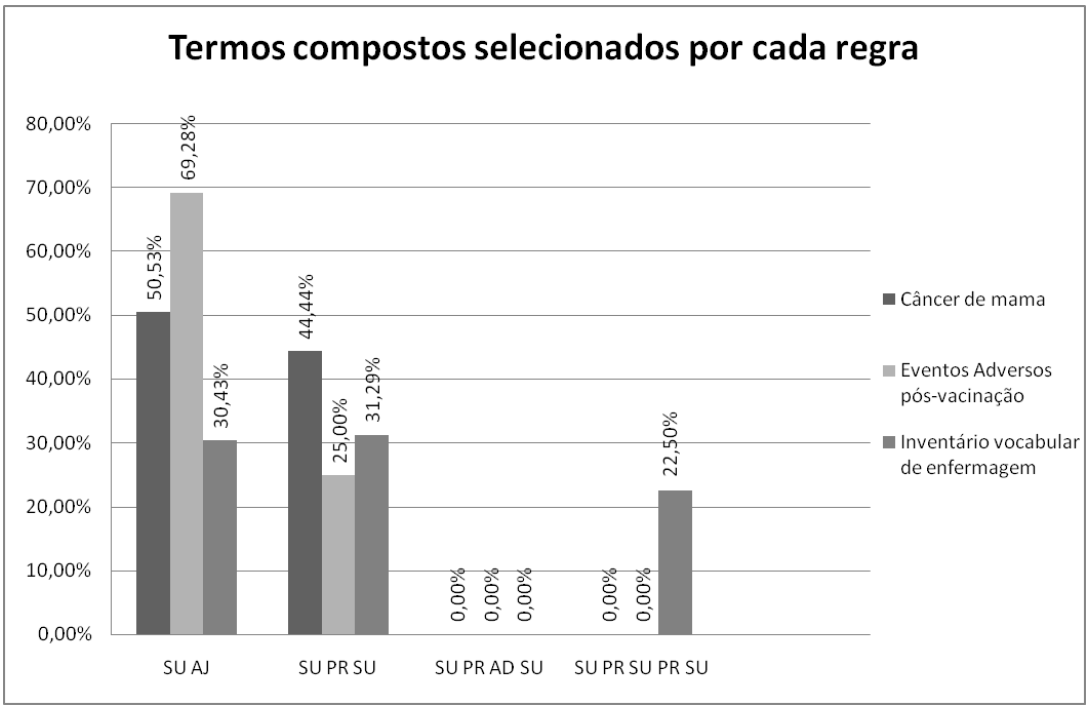


Gráfico 5: Comparação dos termos compostos selecionados por cada regra nos três experimentos
Fonte: Autor

5 CONCLUSÕES

Este trabalho propôs o desenvolvimento de uma ferramenta semiautomática para construção de ontologias a partir de textos em português na área da saúde. A ferramenta proposta foi desenvolvida tendo como base a ferramenta para construção semiautomática de ontologias a partir de textos em inglês Text-to-Onto; as medidas frequência, *tfidf* e entropia; e a ferramenta de anotação linguística TreeTagger.

Foram realizados três experimentos principais para validar a ferramenta, sendo os experimentos nas áreas de domínio do câncer de mama; de eventos adversos pós-vacinação e de inventário vocabular de enfermagem. Os termos extraídos foram validados por especialistas nas respectivas áreas.

Analisando-se os resultados deste trabalho conclui-se que é possível construir ontologias na área da saúde, de maneira semiautomática, a partir de textos em português, sendo que os objetivos propostos foram alcançados, tendo esta dissertação como principais contribuições:

- ferramenta para a construção semiautomática de ontologias a partir de textos em português desenvolvida e validada;
- avaliação prática de ferramentas disponíveis para construção semiautomática de ontologias;
- avaliação prática de ferramentas disponíveis para anotação linguística;
- construção de três novas estruturas ontológicas na área da saúde.

Foi comprovada a dificuldade em se construir ontologias automaticamente devido à complexidade envolvida no processo de extração de termos, sendo muito importante a participação do especialista na etapa do pós-processamento.

É possível também concluir que, dado que a avaliação do especialista é subjetiva, baseada em um domínio e escopo estabelecidos, é preciso selecionar especialistas com critérios bem definidos e padronizados e em quantidade significativa para se ter uma avaliação menos sujeita a falhas, tendo consequentemente uma ontologia de consenso.

Os resultados foram analisados e pode-se concluir com os experimentos que:

- a quantidade de termos extraídos pela ferramenta é dependente da poda realizada pelos filtros que eliminam termos simples e compostos de acordo com a frequência estabelecida;
- a alta frequência com que um termo aparece no corpus não significa que ele será selecionado. Muitos dos termos extraídos de maneira automática pela ferramenta não foram selecionados como relevantes pelos especialistas devido à quantidade de "ruído" no *corpus*;
- é importante incluir outros tipos de filtros, como por exemplo, para exclusão de nomes próprios do *corpus*, sendo possível assim refinar ainda mais a extração dos termos;
- o uso do DeCS como medida para extração pode melhorar os resultados obtidos.

Este trabalho é uma contribuição para o processo de aprendizagem de ontologias, fornecendo uma ferramenta com funcionalidades até então não propostas: não há necessidade do usuário realizar previamente a anotação linguística do *corpus*, estando este acoplado à ferramenta, facilitando sua usabilidade; a disponibilidade da ferramenta em um ambiente *Web*, onde o acesso é facilitado para profissionais e para a comunidade cientifica; a inserção de sinônimos, facilitando e auxiliando na seleção dos termos; e a inserção de termos relacionados com os DeCS, facilitando também a seleção dos termos pelo especialista.

Cabe salientar que a ferramenta desenvolvida poderá ser utilizada em qualquer área de domínio e não apenas na área da saúde. A partir deste estudo, surgem perspectivas de trabalhos futuros, entre as quais:

- melhorar a poda do *corpus* retirando: nomes próprios, referências, quadros, tabelas e outros que possam gerar ruído no processamento;
- utilizar o DeCS como medida de extração;
- criar um filtro com limiar de corte com o máximo de vezes em que um termo pode aparecer em um corpus;
- implementar novas regras de identificação de termos compostos;
- implementar novos métodos para geração taxonômica;
- implementar a extração de relacionamentos (não hierárquicos)
- validar a ferramenta com um *corpus* de outra área.

REFERÊNCIAS

ALFONSECA, E.; RODRÍGUEZ P. **Automatically Generating Hypermedia Documents depending on User Goals**, Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems, Málaga, Spain, 2002.

ALVES, C. F. ; MALUCELLI, A. **Ontologia na área de histocompatibilidade no transplante de órgãos**. In: Seminário de Pesquisa em Ontologia, 2008, Niterói. SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008.

APACHE POI. **Java API To Access Microsoft Format Files.** Disponível em: <http://poi.apache.org>. Acesso em: 15 out. 2009.

AUSSENAC-GILLES, N. ; SEGUELA P. **Les relations sémantiques: du linguistique au formel Cahiers de grammaire**. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : Presse de l'UTM. Pp 175-198, 2000.

BACHIMONT, B. **Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser**. PhD Thesis, Ecole Polytechnique, 1996.

BASÉGIO, T. **Uma Abordagem Semiautomática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil.** Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, 2008.

BIÉBOW, B.; SZULMAN S. **TERMINAE: a linguistic-based tool for the building of a domain ontology**. In EKAW'99 Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, pages 49-66, Berlin, 1999.

BISSON, G.; NEDELLEC, C.; CAÑAMERO, D. **Designing Clustering Methods for Ontology Building. The Mo'K Workbench**. In S. Staab, A. Maedche, C. Nedellec, P. WiemerHasting (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00, Berlin, Germany, 2000.

BORST, W.N. **Construction of engineering ontologies**, in University of Twente. University of Twente: Enschede, 1997.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. **Ontology learning from text: An overview**. In: P-BUITELAAR; CIMIANO, P.; MAGNINI, B. (Ed.). Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, 2005.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. **Ontolt: A protégé plug-in for ontology extraction from text**. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2003.

CAMPOS, M. L. A.; **Integração de Ontologias: o domínio da Bioinformática.** Revista eletrônica de comunicação e inovação em saúde, 2007. Disponível em: < http://bvssp.icict.fiocruz.br/lildbi/docsonline/4/9/994-Campos.pdf>. Acesso em: 5 nov. 2009.

CHAELANDAR, G.; GRAU B. **SVETLAN'- A System to Classify Words in Context**. In S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.) Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, 2000.

CIMIANO, P. **Ontology Learning and Population: Algorithms, Evaluation and Applications**. PhD thesis, University of Karlsruhe, 2005.

CIMIANO, P.; VÖLKER J. **A framework for ontology learning and data-driven change discovery**. In *Proc. of the NLDB'2005*, 2005.

CIRAMITA, M.; GANGEMI, A.; RATSCH, E.; SARIC, J., ROJAS, I. **Unsupervised learning of semantic relations between concepts of a molecular biology ontology**. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005.

DeCS. **DeCS - Descritores em Ciências da Saúde.** Disponível em: em <http://decs.bvs.br/P/DeCS2008_Alfab_por.htm>. Acesso em: 15 out. 2009.

DELLSCHAFT, K.; STAAB, S. **Strategies for the Evaluation of Ontology Learning**, IOS Press, 2007.

ENGELS, R. **CORPORUM OntoExtract. Ontology Extraction Tool**. Deliverable 6 Ontoknowledge. In: Proceedings of the Demo Session of the International, 2001.

EUROWORDNET. **A multilingual database with lexical semantic networks.** Disponível em: <http://www.hum.uva.nl/~ewn>. Acesso em: 22 ago. 2008.

FAURE, D.; POIBEAU T. **First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX.** In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer- Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, 2000.

FAURE, D.; NEDELLEC, C. **Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium.** In: EKAW '99: Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management. London, UK: Springer-Verlag, p. 329-334. ISBN 3-540-66044-5, 1999.

FREITAS, F.; SCHULZ, S. **Ontologias, Web semântica e saúde.** Revista eletrônica de comunicação e inovação em saúde, 2009. Disponível em: <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/238/246>. Acesso em: 5 nov. 2009.

GARCIA, T. R.; NÓBREGA, M. M. L.; SOUSA, M. C. M. **Validação das definições de termos identificados no projeto CIPESC para o eixo foco da prática em enfermagem da CIPE®.** Revista Brasileira de Enfermagem, v. 55, n. 1, p. 52- 63, 2002.

GARCIA, T. R.; NÓBREGA, M. M. L. **Inventário vocabular de fenômenos e ações de enfermagem.** In: Telma Ribeiro Garcia; Maria Miriam Lima da Nóbrega. (Org.). Sistemas de Classificação em Enfermagem: um trabalho coletivo. João Pessoa - PB: Idéias, 2000, v. 1, p. 83-170.

GO. **Gene Ontology.** Disponível em: <http://www.geneontology.org>. Acesso em: 5 nov. 2009.

GÓMEZ-PÉREZ, A.; MANZANO-MACHO D. **A survey of ontology learning methods and Techniques.** Universidad Politécnica de Madrid, 2003.

GRUBER, T.R., **A translation approach to portable ontology specifications. Knowledge Acquisition**. 5: p. 199-220, 1993.

GRÜNINGER, M.; FOX M. S. **Methodology for the design and evaluation of ontologies**, In: IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Canada, 1995.

GUARINO, N. **Formal ontology and information systems.** In: GUARINO, N. (Ed.). Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam: IOS Press. p.3-15, 1998.

GUPTA, K.M., AHA, D.W., MARSH, E., AND MANEY, T. **An architecture for engineering sublanguage WordNets.** In Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages, 2002.

HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora.** In: Proceedings of the 14th conference on Computational linguistics. Morristown, NJ, USA: Association for Computational Linguistics. p. 539-545, 1992.

HUHNS, M. N., SINGH, M. P. **Readings in Agents.** Morgan Kaufmann Publishers, INC. San Francisco, California, 1997.

ICNP®. **International Classification for Nursing Practice.** Disponível em: <http://www.icn.ch/icnpcountdownV1-2.htm>. Acesso em: 5 nov. 2009.

IMGT. **ImMunoGeneTics.** Disponível em: <http://imgt.cines.fr/textes/IMGTindex/ontology.html>. Acesso em: 5 nov. 2009.

JONES, S., PAYNTER, G.W. **Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications.** Journal of the American Society for Information Science and Technology (JASIST), 2002;

LIN, D.; PANTEL, P. **Dirt - discovery of inference rules from text.** In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323–328, 2001.

LINGUATECA. **Um centro de recursos distribuído para o processamento computacional da língua portuguesa.** Disponível em: <http://www.linguateca.pt>. Acesso em: 01 set. 2009.

MAEDCHE, A.; STAAB, S. **Ontology Learning.** In S. Staab & R. Studer (eds.) Handbook on Ontologies in Information Systems, 2003. <http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf>. Acesso em: 3 nov. 2009.

MAEDCHE, A.; STAAB, S. **Ontology learning for the semantic *web*.** IEEE Intelligent Systems, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 2, p. 72-79, 2002.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing.** Cambridge, Massachusetts: The MIT Press, 1999.

MHC. **Major Histocompatibility Complex.** Disponível em: <http://www.jimmunol.org/cgi/content/meeting_abstract/178/MeetingAbstracts/LB>. Acesso em: 5 nov. 2009.

MICROSOFT OFFICE. Disponível em: <http://office.microsoft.com/pt-br/default.aspx>. Acesso em: 15 out. 2009.

MIKHEEV, A., FINCH, S. **A Workbench for Finding Structure in Texts.** Proceedings of ANLP-97 (Washington D.C.), 1997.

MINISTÉRIO DA SAÚDE. **Manual de vigilância epidemiológica de eventos adversos pós-vacinação / Ministério da Saúde, Secretaria de Vigilância em Saúde**, Departamento de Vigilância Epidemiológica. – Brasília: Ministério da Saúde, 2008.

MORIN, E. **Acquisition de patrons lexico-syntaxiques caractéristiques dúne relation sémantique.** TAL (Traitement Automatique des Langues). 40/1: 143-166, 1999.

MORIN, E.; JACQUEMIN, C. **Automatic acquisition and expansion of hypernym links. Computers and the Humanities**, v. 38, n. 4, p. 363-396, 2004.

MUSEN, M.A. **Medical informatics: searching for underlying components**. Methods Inf Med, 2002.

MYFACES. **Apache MyFaces - project of the Apache Software Foundation.** Disponível em: <http://myfaces.apache.org >. Acesso em: 01 set. 2009.

NECHES, R.; FIKES, R. E.; FININ, T.; GRUBER, T. R., SENATOR, T.; SWARTOUT, W. R. **Enabling technology for knowledge sharing**, AI Magazine, Vol.12(3):36-56, 1991.

OBO. **Open Biomedical Ontologies.** Disponível em: <http://www.obofoundry.org>. Acesso em: 5 nov. 2009.

OPENTHESAURUSPT. **Um projeto Open Source para a construção de um Dicionário de Sinônimos para a língua portuguesa.** Disponível em: em <http://openthesaurus.caixamagica.pt >. Acesso em: 15 out. 2009.

PDFBOX. **Apache PDFBox - Java PDF Library.** Disponível em: <http://incubator.apache.org/pdfbox >. Acesso em: 15 out. 2009.

RIBEIRO JUNIOR, L. C. **OntoLP: Construção semiautomática de ontologias a partir de Textos da Língua Portuguesa.** Dissertação (Mestrado) - Universidade do Vale do Rio dos Sinos- UNISINOS, 2008.

RICHFACES. **JBoss Richfaces component library.** Disponível em: <http://www.jboss.org/jbossrichfaces >. Acesso em: 01 set. 2009.

RYU, P.-M.; CHOI, K. S. **Taxonomy learning using term specificity and similarity.** In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Sydney, Australia: Association for Computational Linguistics, 2006. p. 41-48.Semantic *Web* Conference (ISWC). Sanibel Island, Florida,: [s.n.], 2003.

SCHMID, H. **TreeTagger - a language independent part-of-speech tagger,** 1994. Disponível em: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>. Acesso em: 03 nov. 2009.

SILVA, R. R.; MALUCELLI, A.; CUBAS, M. R. **Em Direção à Ontologia CIPESC®.** In: XI Congresso Brasileiro de Informática em Saúde, 2008, Campos do Jordão. XI Congresso Brasileiro de Informática em Saúde, 2008.

SMITH, M. K.; WELTY, C.; MCGUINNESS, D. L. **OWL Web Ontology Language Guide**, 2004. Disponível em: <http://www.w3.org/TR/owl-guide/>. Acesso em: 05 nov. 2009.

SNOMED. **Systemized Nomenclature of MEDicine.** Disponível em: <http://www.snomed.org>. Acesso em: 5 nov. 2009.

STUDER, R.; BENJAMINS, R.; FENSEL, D. **Knowledge Engineering: Principles and Methods.** Data and Knowledge Engineering. 25(1998):161-197, 1998.

THOMPSON, C.A.; MOONEY, R. J. **Semantic Lexicon Acquisition for Learning Parsers.** Technical Note, 1997.

UMLS. **Unified Medical Language System.** Disponível em: <http://www.nlm.nih.gov/research/umls>. Acesso em: 5 nov. 2009.

USCHOLD, M.; GRÜNINGER, M. **Ontologies: Principles,Methods and Applications**, Knowledge Engineering Review, Vol.11(2):93-155, 1996.

USCHOLD, M.; JASPER, R**. A Framework for Understanding and Classifying Ontology Applications**, In: Benjamins VR (ed) IJCAI'99 Workshop on Ontology and Problem Solving Methods: Lessons Learned and Future Trends, Stockholm, Sweden, CEUR Workshop Proceedings 18:11.1-11.12. Amsterdam, The Netherlands, 1999.

van HEIJST, G.; SCHREIBER, A. Th.; WIELINGA, B. J. **Using explicit ontologies in KBS development**, International Journal of Human-Computer Studies, Vol.45:184-292, 1997.

VELARDI, P.; NAVIGLI R.; AND MISSIKOFF M. **Integrated approach for *Web* ontology learning and engineering**. IEEE Computer, 2002.

VISL. **Visual Interactive Syntax Learning.** Disponível em: <http://visl.hum.sdu.dk/visl/pt>. Acesso em: 15 out. 2009.

WEISS, G. **Multiagent systems: a modern approach to distributed artificial intelligence**, The MIT Press Cambridge, Massachusetts, London, England, 1999.

WIENER, N. **Cybernetics.** In: FOERSTER, H. V. Cybernetics of Cybernetics: the control of control and the communication of communication. Minneapolis: Future Systems, p. 7-17, 1948.

WORDNET. **An electronic lexical database.** Disponível em: <http://wordnet.princeton.edu>. Acesso em: 22 ago. 2008.

WU, S.H.; HSU, W.L. **SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese *Corpus*.** In the 19th International Conference on Computational Linguistics, Howard International House and Academia Sinica, Taipei, Taiwan, 2002.

XU, F.; KURZ, D.; PISKORSKI, J.; SCHMEIER, S. **A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping.** In Proceedings of LREC 2002, the third international conference on language resources and evaluation, Las Palmas, Canary island, Spain, 2002.

APÊNDICE I

Resultado no experimento da área de câncer de mama									
Termo	**Total**	**Tf-Idf**	**Entropy**	**Decs**	**Termo**	**Total**	**Tf-Idf**	**Entropy**	**Decs**
abordagem	93	8,02	1,19	F	linfonodos regionais	176	8,89	1,02	F
ação	788	7,51	1,19	F	líquido	52	8,31	1,19	F
acesso	152	7,73	1,23	F	local	36	8,11	1,2	F
ações de controle	31	8,67	1,17	F	localização	2178	7,73	1,18	V
ações de enfermagem	193	8,96	1,03	F	localização primária	304	8,31	1,07	F
acompanhamento	83	8,02	1,2	F	mãe	33	8,56	1,16	F
adjuvante	46	9,12	1,11	F	mal	199	8,71	1,09	F
administração	141	8,2	1,17	V	malig	85	8,71	1,04	F
adulto	47	8,11	1,21	V	malig de cel	40	10,75	1	F
agulha	32	8,11	1,07	F	malignant	35	8,31	1,17	F
ajustada	70	10,5	1	F	malignidade	48	7,94	1,21	F
ajuste	32	8,56	1,13	F	maligno	662	7,61	1,21	F
álcool	69	8,42	1,12	F	mama	1157	7,41	1,24	V
alimentação	73	8,02	1,16	V	mama feminina	33	8,26	1,22	F
alimento	76	7,94	1,15	V	mamografia	227	8,11	1,11	V
alívio	71	9,12	1,09	F	management	34	8,42	1,17	F
alta	126	7,86	1,22	F	margem	49	8,71	1,15	V
alteração	176	7,67	1,2	F	massa	51	8,02	1,2	F
âmbito	73	8,02	1,19	F	mastologia	48	8,71	1,13	F
ambulatório	30	8,2	1,18	F	mau	63	9,81	1,01	F
amostra	35	8,42	1,14	F	medicação	84	8,71	1,11	F
ampliação	57	8,11	1,2	F	medicamento	126	7,86	1,17	F
analgesia	67	9,12	1,08	V	medicina	119	7,94	1,19	V
analgésico	76	8,71	1,1	F	médico	327	7,73	1,21	F
análise	229	7,56	1,24	F	medo	34	8,89	1,11	V
anatomia	71	8,42	1,13	V	menor	109	8,11	1,17	F
anatomia patológica	39	9,12	1,09	F	meta	39	8,2	1,19	F
anual	211	8,71	1,14	F	metástase	634	8,31	1,09	V
aplicação	119	8,2	1,16	F	metástase em linfonodos	103	10,75	1	F
assistência	377	7,61	1,22	V	método	155	7,51	1,24	F
assistência oncológica	76	8,42	1,12	F	metodologia	69	7,94	1,22	V
associado	55	7,8	1,23	F	microcalcificações	41	9,12	1,07	F
atenção oncológica	38	8,31	1,19	F	ministério	69	9,81	1,13	F
atendimento	192	7,8	1,21	F	monitoramento	59	8,02	1,16	V
atualização	63	7,94	1,19	F	morbidade	50	7,73	1,25	V
aumento	223	7,46	1,25	F	morfina	128	9,4	1,08	V

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
ausência de metástase	93	10,06	1,01	F	mortalidade	829	7,61	1,21	V
avançado	31	8,02	1,22	F	mortalidade de número	66	9,4	1,07	F
base populacional	37	8,02	1,22	F	mortalidade por câncer	151	7,92	1,23	F
benefício	54	8,02	1,21	F	mortalidade por casos	66	8,11	1,19	F
bioética	32	8,31	1,11	V	mortality	74	8,02	1,15	F
biópsia	104	8,2	1,19	V	morte	160	7,73	1,21	V
brasileiro de geografia	194	8,67	1,06	F	movimento	64	8,02	1,2	V
breast	56	8,31	1,19	F	ms	180	7,67	1,18	F
brutas de incidência	540	8,96	1,14	F	msincaconprevdivisão	140	9,81	1,03	F
busca	46	8,11	1,21	F	mulher	1389	7,67	1,23	F
ca	54	8,02	1,17	F	mulheresd	32	10,5	1	F
cacon	99	8,31	1,19	F	músculo	68	8,31	1,17	F
cálculo	57	7,67	1,25	F	mutação	34	8,42	1,17	V
cân	30	8,2	1,2	F	mx	42	9,12	1,04	F
cancer	535	7,51	1,22	F	náusea	50	8,89	1,12	V
câncer	5909	7,41	1,26	V	neop	240	10,5	1	F
câncer de mama	31	8,56	1,1	V	neop malig	40	10,5	1	F
câncer em mulheres	32	8,02	1,14	F	neoplasia	1535	7,51	1,2	F
cancerologia	138	8,11	1,13	V	neoplasia maligna	93	8,35	1,1	F
capacidade	81	8,02	1,21	F	neoplasias de células	41	10,75	1	F
capacitação	66	8,02	1,19	V	neoplasias de tecido	32	10,75	1	F
captação	49	8,56	1,16	F	neoplasias malignas	80	8,71	1,1	F
carcinoma	760	7,86	1,15	V	nível	280	7,46	1,24	F
caso	2575	7,51	1,23	F	nódulo	170	8,42	1,14	F
casos de câncer	69	8,04	1,23	F	norma	79	8,02	1,2	F
casos novos	312	8,02	1,17	F	novo	958	8,11	1,16	F
casos por tipo	52	10,06	1,06	F	núcleo	40	8,02	1,21	F
causa	229	7,73	1,17	V	nutrição	100	8,71	1,08	V
causas de morte	35	8,96	1,13	V	nutrição oncológica	55	10,5	1	F
célula	548	7,73	1,18	F	óbito	588	7,86	1,13	F
celulas	40	10,5	1	F	óbitos por câncer	166	8,8	1,05	F
chemotherapy	49	9,12	1,1	F	ocorrência	102	7,86	1,23	F
científica	172	7,73	1,11	F	oncol	41	9,12	1,05	F
cigarro	64	8,2	1,16	V	oncologia	351	7,67	1,18	V
cirurgia	365	7,61	1,19	V	oncológicos	153	9,81	1,07	F
cirurgião	47	8,56	1,1	F	oncologista	36	8,42	1,14	F
citologia	30	8,42	1,18	V	oncology	69	8,31	1,17	F
citopatologia	38	8,89	1,13	F	operacional	36	8,89	1,03	F
classificação	496	7,61	1,13	V	orçamento	44	8,2	1,2	F
classificação aplicável	32	10,5	1	F	organismo	33	8,71	1,09	F
classification	67	8,89	1,1	F	óssea	43	8,71	1,16	F
clínica	280	7,67	1,21	F	osso	82	8,89	1,14	V

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
clinical	33	8,31	1,18	F	ovário	32	8,2	1,16	V
clínico	249	7,61	1,18	F	padronização	35	8,2	1,2	F
cobertura	161	7,86	1,19	F	paliativo	302	7,86	1,17	F
codeína	36	9,4	1,07	V	palpação	47	8,71	1,11	V
coleta	69	8,02	1,19	F	parede	98	8,31	1,13	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
familiar	71	8,2	1,18	F	seio	151	9,4	1,07	F
fase de intensificação	49	9,65	1,06	F	sensibilidade	40	8,42	1,14	F
fatores de risco	180	7,75	1,22	V	sentinela	34	8,71	1,1	F
females	38	10,5	1	F	serviços de saúde	78	8,35	1,18	V
feminino	400	7,67	1,18	V	sexo	530	7,8	1,15	V
fibrossarcomas	41	10,5	1	F	sexo feminino	87	7,94	1,17	F
filho	68	7,86	1,21	F	sinal	113	8,31	1,14	F
financeiro	31	8,42	1,19	F	sintoma	216	8,11	1,17	F
fisiopatologia	33	9,81	1,03	F	sobrevida	272	7,67	1,18	V
fisioterapia	62	8,71	1,08	F	sobrevida média	31	9,12	1,14	F
fluxo	33	8,2	1,19	F	social	44	8,31	1,05	F
formação	174	7,8	1,19	F	substituição	30	8,56	1,14	F
freqüência	151	8,11	1,17	F	sus	235	7,73	1,24	V
fumante	53	8,56	1,15	F	suspeita	45	8,56	1,12	F
fumo	49	8,2	1,18	V	tabaco	157	8,56	1,16	V
funcionamento	31	8,2	1,18	F	tabagismo	250	7,67	1,19	V
genética	39	8,31	1,18	V	tamanho	59	8,56	1,16	F
ginecologia	43	8,31	1,12	V	taxa	2533	7,51	1,2	F
glândula	417	8,56	1,05	F	taxa de mortalidade	30	8,35	1,19	V
gordura	55	8,02	1,2	F	taxa específica	51	9,81	1,06	F
grupamento por estádios	67	10,75	1	F	taxas bruta	81	9,81	1,01	F
grupos de casos	41	10,06	1,02	F	taxas brutas	200	8,96	1,14	F
histológico	31	10,5	1	F	taxas de incidência	193	8,26	1,16	F
histopatológica	49	10,5	1	F	taxas de mortalidade	109	8,26	1,15	F
hormônio	34	8,71	1,14	F	taxas médias	30	9,81	1,07	F
hospitalar	59	8,2	1,18	F	taxas por milhão	70	10,75	1	F
iarc	89	7,86	1,2	F	tecido linfático	32	10,75	1	F
identificação	56	8,02	1,2	F	tecido mole	36	9,81	1,02	F
imagem	234	7,94	1,15	F	tecidos moles	87	9,12	1,03	F
impacto	67	7,56	1,24	F	técnica	186	7,67	1,25	F
importância	128	7,73	1,2	F	tecnologia	60	7,94	1,2	V
inca	984	7,51	1,2	F	tendência	117	7,61	1,23	F
incams	66	8,02	1,19	F	terapêutica	66	8,11	1,19	V
incidence	58	8,56	1,1	F	terapêutico	50	8,2	1,2	F
incidência	1646	7,56	1,22	V	terapia	122	7,61	1,21	F
incidência brutas	64	10,75	1	F	therapy	45	8,56	1,13	F
incidência de câncer	89	8,18	1,21	F	tipo celular	94	10,5	1	F
incidência por casos	128	8,56	1,2	F	tipo de câncer	216	8,55	1,1	F
incidência por mulheres	96	9,36	1,13	F	tipos de câncer	76	8,26	1,2	F
incidência por tipo	118	10,75	1	F	tnm	136	8,89	1,05	F
incorporação	30	8,2	1,19	F	total	692	7,46	1,2	F
indicação	89	8,2	1,19	F	tratamento	946	7,51	1,21	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
indicador	111	7,67	1,23	F	tratamento cirúrgico	34	8,71	1,14	F
infra-estrutura	42	8,89	1,1	F	tratamento de câncer	34	8,67	1,15	F
inspeção	31	8,89	1,1	F	treatment	54	8,42	1,15	F
instituto de câncer	660	7,75	1,22	F	treinamento	84	8,2	1,17	F
insuficiência	35	8,89	1,13	F	tu	122	9,81	1,01	F
intensiva	39	8,89	1,1	F	tu de cel	80	10,75	1	F
internação	87	8,2	1,19	F	tumor	3156	7,46	1,18	F
internet	54	7,94	1,22	V	tumor por faixa	128	10,75	1	F
intervenção	82	8,11	1,19	F	tumor primário	41	9,36	1,06	F
invasão	132	8,56	1,08	F	tumor residual	51	10,5	1	V
investimento	43	8,2	1,2	F	tumores de células	54	9,65	1,02	F
junção	71	8,71	1,09	F	tumores malignos	213	8,02	1,13	F
laboratórios de anatomia	40	9,36	1,09	F	tumores ósseos	142	8,56	1,04	F
laudo	33	8,42	1,18	F	tumores primários	52	9,65	1,04	F
leito	78	8,31	1,18	F	tumors	33	8,89	1,12	F
lesão	342	7,61	1,2	F	tx	47	9,81	1,01	F
linfadenectomia	32	9,12	1,09	V	ulceração	38	8,42	1,14	F
linfático	39	8,56	1,16	F	ultrasonografia	53	8,71	1,14	V
linfonodo	43	10,06	1,04	F	unidades de saúde	84	8,67	1,16	F
linfonodos	581	8,56	1,05	V	vigilância	257	7,56	1,26	V
linfonodos axilares	50	8,89	1,1	F					

Quadro 19: Resultado no experimento na área de câncer de mama

APÊNDICE II

Resultado do experimento da área de eventos adversos pós-vacinação									
Termo	**Total**	**Tf-Idf**	**Entropy**	**Decs**	**Termo**	**Total**	**Tf-Idf**	**Entropy**	**Decs**
abscesso	42	8,46	1	V	icterícia	9	8,46	1	V
abscesso quente	4	8,46	1	F	igm	5	8,46	1	V
acompanhamento	12	8,46	1	F	im	9	8,46	1	F
acompanhamento neurológico	3	8,46	1	F	imagem	3	8,46	1	F
adjuvante	11	8,46	1	F	imunidade	11	8,46	1	V
administração	51	8,46	1	V	imunização	52	8,46	1	V
adrenalina	4	8,46	1	V	imunização passiva	3	8,46	1	V
agente	5	8,46	1	F	imunobiológicos	13	8,46	1	F
agudo	6	8,46	1	F	imunodeficiência	9	8,46	1	F
alergia	19	8,46	1	V	imunodepressão	9	8,46	1	F
alumínio	11	8,46	1	V	imunodeprimidos	3	8,46	1	F
amostra	28	8,46	1	F	imunoenzimático	3	8,46	1	F
anafilaxia	34	8,46	1	V	imunoglobulina	14	8,46	1	F
analgésico	5	8,46	1	F	imunoglobulinas humanas	4	8,46	1	F
angioedema	5	8,46	1	V	imunohistoquímica	5	8,46	1	F
anticorpo	15	8,46	1	F	inativação	3	8,46	1	F
antígeno	18	8,46	1	F	incidência	26	8,46	1	V
antitérmico	4	8,46	1	F	incqs	4	8,46	1	F
antitoxina	4	8,46	1	F	infantil	3	8,46	1	F
aplicação	120	8,46	1	F	infecção	30	8,46	1	V
aplicação intramuscular	3	8,46	1	F	inflamação	7	8,46	1	V
aplicação simultânea	3	8,46	1	F	influenza	19	8,46	1	F
apnéia	7	8,46	1	V	influenzae	7	8,46	1	F
arthus	12	8,46	1	F	informação	28	8,46	1	V
artralgia	7	8,46	1	V	injeção	16	8,46	1	F
artrite	11	8,46	1	V	inoculação	9	8,46	1	F
assistência	7	8,46	1	V	instalação súbita	3	8,46	1	F
associações temporais	5	8,46	1	F	intensidade leve	3	8,46	1	F
associado	20	8,46	1	F	intervalo	10	8,46	1	F
associados com infecção	3	8,46	1	F	intussusception	3	8,46	1	F
astenia	8	8,46	1	V	invaginação	10	8,46	1	V
ataxia	9	8,46	1	V	invaginação intestinal	4	8,46	1	V
atrofia	4	8,46	1	V	investigação	119	8,46	1	F
atrofia muscular	3	8,46	1	V	investigação neurológica	4	8,46	1	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
autismo	3	8,46	1	F	irritabilidade	16	8,46	1	F
avaliação	48	8,46	1	V	irritação	5	8,46	1	F
bacilo	5	8,46	1	F	isolamento	9	8,46	1	F
bcg	24	8,46	1	V	isolamento viral	4	8,46	1	F
bcgid	3	8,46	1	F	isoniazida	11	8,46	1	V
biologia molecular	4	8,46	1	V	lactente	3	8,46	1	V
broncoespasmo	3	8,46	1	V	laringoespasmo	9	8,46	1	V
calendário básico	3	8,46	1	F	lcr	7	8,46	1	F
calor	5	8,46	1	V	leite materno	4	8,46	1	V
caso	205	8,46	1	F	leningradzagreb	3	8,46	1	F
casos graves	17	8,46	1	F	lesão	36	8,46	1	F
causalidade	3	8,46	1	V	lesão vacinal	3	8,46	1	F
caxumba	11	8,46	1	V	lesões locais	3	8,46	1	F
cefaléia	28	8,46	1	V	leucocitose	7	8,46	1	V
cefaléia intensa	3	8,46	1	F	leucopenia	7	8,46	1	V
células diplóides	3	8,46	1	F	linfadenopatia	9	8,46	1	V
cepa	43	8,46	1	F	linfadenopatia regional	7	8,46	1	F
choque	29	8,46	1	V	linfonodos	7	8,46	1	V
choque anafilático	9	8,46	1	V	local	60	8,46	1	F
choro	6	8,46	1	V	lote	15	8,46	1	F
cianose	9	8,46	1	V	manifestações alérgicas	7	8,46	1	F
cicatrização	7	8,46	1	F	manifestações locais	8	8,46	1	F
cobertura	7	8,46	1	F	manifestações neurológicas	11	8,46	1	V
coberturas vacinais	3	8,46	1	F	manifestações sistêmicas	14	8,46	1	F
complemento	8	8,46	1	F	marcha	3	8,46	1	V
complicação	19	8,46	1	F	medula óssea	3	8,46	1	V
componente	40	8,46	1	F	meningite	19	8,46	1	V
componente da rubéola	4	8,46	1	F	meningite asséptica	4	8,46	1	V
componente da vacina	4	8,46	1	F	meningococo	15	8,46	1	F
componentes da vacina	9	8,46	1	F	mialgia	11	8,46	1	F
conduta	101	8,46	1	V	mielite	3	8,46	1	V
contaminação	8	8,46	1	V	ml	11	8,46	1	F
contato	18	8,46	1	F	monitoramento	3	8,46	1	V
contraindicação	59	8,46	1	F	morte súbita	3	8,46	1	V
contraindicacões	4	8,46	1	F	mutação	4	8,46	1	V
convulsão	39	8,46	1	F	náusea	10	8,46	1	V
convulsão febril	5	8,46	1	F	neurite	8	8,46	1	V
convulsões febris	4	8,46	1	V	neurovirulência	5	8,46	1	F
coqueluche	11	8,46	1	V	nódulo	4	8,46	1	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
corticosteróide	11	8,46	1	F	notificação	93	8,46	1	V
crie	13	8,46	1	F	notificar	82	8,46	1	F
crise convulsiva	3	8,46	1	F	óbito	20	8,46	1	F
cuidado	3	8,46	1	F	observação	15	8,46	1	V
cultivo	13	8,46	1	F	ºc	16	8,46	1	F
cultivo celular	3	8,46	1	F	palidez	11	8,46	1	V
dado	34	8,46	1	F	pápula	8	8,46	1	F
dados clínicos	3	8,46	1	F	paracetamol	3	8,46	1	V
deficiência imunológica	3	8,46	1	F	paralisia	20	8,46	1	V
déficit	17	8,46	1	F	paralisia flácida	3	8,46	1	F
diagnóstico	56	8,46	1	V	parotidite	5	8,46	1	V
diagnóstico diferencial	13	8,46	1	V	pcr	8	8,46	1	V
diagnóstico etiológico	9	8,46	1	F	pneumococo	12	8,46	1	F
diagnósticos diferenciais	5	8,46	1	F	poliomielite	28	8,46	1	V
diarréia	7	8,46	1	V	poliovírus	10	8,46	1	F
difteria	9	8,46	1	V	polirradiculoneurite	3	8,46	1	V
diluente	3	8,46	1	F	polissacarídeo	12	8,46	1	F
diluição	3	8,46	1	V	polissacarídeos capsulares	3	8,46	1	F
disfunção	6	8,46	1	F	portador	3	8,46	1	V
dispnéia	5	8,46	1	V	potência	3	8,46	1	V
disseminação	6	8,46	1	F	precaução	31	8,46	1	V
dissociação albuminocitológica	3	8,46	1	F	prevenção	3	8,46	1	F
doença	109	8,46	1	V	primovacinados	8	8,46	1	F
doença do soro	7	8,46	1	V	prognóstico	8	8,46	1	V
doença infecciosa	3	8,46	1	F	proteína	13	8,46	1	F
doença neurológica	5	8,46	1	F	protocolo	8	8,46	1	F
doença viscerotrópica	4	8,46	1	F	púrpura	19	8,46	1	F
doenças agudas	8	8,46	1	F	púrpura pósvacinal	3	8,46	1	F
doenças neurológicas	3	8,46	1	F	púrpura trombocitopênica	11	8,46	1	V
dor	66	8,46	1	V	pústula	4	8,46	1	F
dor da injeção	3	8,46	1	F	raiva	10	8,46	1	V
dose	164	8,46	1	V	reação	85	8,46	1	F
dose máxima	3	8,46	1	F	reação alérgica	4	8,46	1	F
doses subseqüentes	41	8,46	1	F	reação anafilática	21	8,46	1	V
dt	19	8,46	1	F	reação grave	4	8,46	1	F
dtp	47	8,46	1	F	reação sistêmica	3	8,46	1	F
dtp acelular	6	8,46	1	F	reações adversas	3	8,46	1	F
dtp celular	5	8,46	1	F	reações alérgicas	6	8,46	1	F
dtpa	3	8,46	1	V	reações anafiláticas	12	8,46	1	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
dtphib	17	8,46	1	F	reações locais	10	8,46	1	F
dupla bacteriana	4	8,46	1	F	reatogenicidade	3	8,46	1	F
dupla viral	5	8,46	1	F	receptor	4	8,46	1	F
eapv	17	8,46	1	F	recombinante	3	8,46	1	F
edema	46	8,46	1	V	recorrência	4	8,46	1	V
eficácia	6	8,46	1	V	reforço	11	8,46	1	F
elevação da temperatura	3	8,46	1	F	revacinação	11	8,46	1	F
elisa	4	8,46	1	V	risco	76	8,46	1	V
emergência	7	8,46	1	F	riscobenefício	6	8,46	1	F
encefalite	18	8,46	1	V	rubéola	18	8,46	1	F
encefalomielite	5	8,46	1	V	rubor	16	8,46	1	V
encefalopatia	17	8,46	1	V	sarampo	19	8,46	1	V
encefalopatia aguda	3	8,46	1	F	saúde	268	8,46	1	V
enduração	24	8,46	1	F	seqüela	17	8,46	1	F
enfartamento	3	8,46	1	V	seqüela neurológica	4	8,46	1	F
eritema	27	8,46	1	V	sgb	13	8,46	1	F
erradicação	6	8,46	1	F	sibilo	11	8,46	1	V
erradicação da poliomielite	3	8,46	1	F	sinal	27	8,46	1	F
erros técnica	6	8,46	1	F	sintoma	50	8,46	1	F
esquema tríplice	3	8,46	1	F	sistêmicos	8	8,46	1	F
esquema vacinal	3	8,46	1	F	sonolência	16	8,46	1	V
evento	457	8,46	1	F	soro	49	8,46	1	V
evento adverso	5	8,46	1	F	sorogrupo	12	8,46	1	F
eventos adversos	315	8,46	1	F	sorologia	7	8,46	1	V
eventos graves	5	8,46	1	F	sorotipo	9	8,46	1	F
eventos locais	3	8,46	1	F	surto	16	8,46	1	F
eventos neurológicos	5	8,46	1	F	suscetibilidade	4	8,46	1	V
evidência	13	8,46	1	F	suspeito	3	8,46	1	F
evolução	19	8,46	1	V	taxa	10	8,46	1	F
evolução benigna	5	8,46	1	F	temperatura	14	8,46	1	V
exame bacteriológico	8	8,46	1	F	temperatura axilar	5	8,46	1	F
exame histopatológico	5	8,46	1	F	tétano	12	8,46	1	V
exame neurológico	3	8,46	1	V	timerosal	16	8,46	1	V
exantema	5	8,46	1	V	título	6	8,46	1	F
exposição	13	8,46	1	F	tontura	5	8,46	1	V
faixa etária	3	8,46	1	F	tônus muscular	5	8,46	1	V
febre	186	8,46	1	V	toxicidade	3	8,46	1	V
febre alta	4	8,46	1	F	toxóide	20	8,46	1	F
febre amarela	34	8,46	1	V	toxóide tetânico	10	8,46	1	V

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
febre baixa	4	8,46	1	F	tratamento	108	8,46	1	F
febre elevada	3	8,46	1	F	tratamento específico	10	8,46	1	F
febre tifóide	11	8,46	1	V	tratamento sintomático	33	8,46	1	F
feto	3	8,46	1	V	tríplice viral	5	8,46	1	F
fistulização	3	8,46	1	F	tuberculose	8	8,46	1	V
formas graves	3	8,46	1	F	tumefação	7	8,46	1	F
fraqueza	8	8,46	1	F	úlcera	10	8,46	1	F
freqüência	71	8,46	1	F	úlcera profunda	3	8,46	1	F
gânglio	6	8,46	1	F	urticária	30	8,46	1	V
gastrenterite	3	8,46	1	F	urticária exantema	4	8,46	1	F
gastrointestinais	3	8,46	1	F	uso subcutâneo	3	8,46	1	F
genética	3	8,46	1	V	vacina	361	8,46	1	F
genoma viral	3	8,46	1	V	vacina oral	5	8,46	1	F
gestação	5	8,46	1	V	vacina pneumocócica	3	8,46	1	F
gestante	3	8,46	1	F	vacina tetravalente	3	8,46	1	F
grave	4	8,46	1	F	vacina tríplice	7	8,46	1	V
grávida	3	8,46	1	F	vacinação	140	8,46	1	V
haemophilus	6	8,46	1	V	vacinação dos primovacinados	5	8,46	1	F
hemorragia	6	8,46	1	V	vacinas virais	3	8,46	1	V
hepatite	27	8,46	1	V	vacinas vivas	6	8,46	1	F
heterólogo	15	8,46	1	F	varicela	7	8,46	1	V
heterólogos	7	8,46	1	F	vasodilatação	3	8,46	1	V
hib	14	8,46	1	V	vermelhidão	14	8,46	1	F
hib conjugada	3	8,46	1	F	vesícula	6	8,46	1	V
hipersensibilidade	61	8,46	1	V	vigilância	400	8,46	1	V
hiporresponsividade	3	8,46	1	F	vigilância dos eventos	8	8,46	1	F
hipotensão	26	8,46	1	V	vigilância epidemiológica	173	8,46	1	V
hipotensão arterial	3	8,46	1	F	vírus	54	8,46	1	V
hipótese	3	8,46	1	F	vírus vacinal	16	8,46	1	F
hipotonia	4	8,46	1	V	vírus vivos	6	8,46	1	F
hipotônicohiporresponsivo	11	8,46	1	F	visceralização	9	8,46	1	F
histamina	6	8,46	1	V	vômito	20	8,46	1	V
histopatologia	5	8,46	1	F	vop	14	8,46	1	F
hospitalização	3	8,46	1	V	vorh	7	8,46	1	F

Quadro 20: Resultado do experimento da área de eventos adversos pós-vacinação

APÊNDICE III

Resultado do experimento da área de inventário vocabular									
Termo	**Total**	**Tf-Idf**	**Entropy**	**Decs**	**Termo**	**Total**	**Tf-Idf**	**Entropy**	**Decs**
aborto	2	6,45	1	V	galeria	2	6,45	1	F
acompanhamento	7	6,45	1	F	galerias abertas	1	6,45	1	F
acompanhamento da pessoa	1	6,45	1	F	gente	60	6,45	1	F
acompanhamento de hipertensos	1	6,45	1	F	germe	3	6,45	1	F
acompanhamento médico	2	6,45	1	F	gestante	5	6,45	1	F
acv	2	6,45	1	F	gestante novinha	1	6,45	1	F
agente	10	6,45	1	F	ginecologista	2	6,45	1	F
agente comunitário	1	6,45	1	F	gravidez	4	6,45	1	V
água	6	6,45	1	F	higiene	3	6,45	1	V
água da rua	1	6,45	1	F	hipoglicemia	2	6,45	1	V
água de esgoto	1	6,45	1	F	idade	5	6,45	1	F
águas pluviais	2	6,45	1	F	importância	5	6,45	1	F
aids	4	6,45	1	V	incentivo	3	6,45	1	F
ajuda	2	6,45	1	F	informação	6	6,45	1	V
alcoólatras anônimos	1	6,45	1	F	injeção de insulina	1	6,45	1	F
alimentação	2	6,45	1	V	insulina	2	6,45	1	V
aplicação de injeção	1	6,45	1	F	irmão	2	6,45	1	F
aplicação de injeção de insulina	1	6,53	1	F	leite	8	6,45	1	V
atenção	2	6,45	1	V	leite em pó	1	6,45	1	V
atividade	6	6,45	1	F	leite materno	3	6,45	1	V
autoridades competentes	1	6,45	1	F	levantamento	3	6,45	1	F
avc	2	6,45	1	V	livro	3	6,45	1	F
banheiro	2	6,45	1	F	lixo	7	6,45	1	V
bebida alcoólica	1	6,45	1	F	lixo de garrafas	1	6,45	1	F
bloqueio	2	6,45	1	F	mãe	15	6,45	1	F
bombom	7	6,45	1	F	mama	2	6,45	1	V
brecha da janela	1	6,45	1	F	mamadeira	2	6,45	1	V
cabeça	3	6,45	1	V	mamografia	3	6,45	1	V

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
cadastro da família	1	6,45	1	F	maneira	4	6,45	1	F
caminho	2	6,45	1	F	medicação	7	6,45	1	F
campanha de prevenção	1	6,45	1	F	médico	16	6,45	1	F
canal aberto	1	6,45	1	F	menina	7	6,45	1	F
câncer	2	6,45	1	V	morte	2	6,45	1	V
câncer do colo do útero	1	6,53	1	V	mulher	11	6,45	1	F
casos de aborto	1	6,45	1	F	município	2	6,45	1	F
cenoura	2	6,45	1	V	neném	2	6,45	1	F
chá	7	6,45	1	V	obrigação	2	6,45	1	F
chá da flor do sabugueiro	1	6,53	1	F	ônibus da prefeitura	1	6,45	1	F
chá da folha do chuchu	1	6,53	1	F	orientação	2	6,45	1	V
chá da folha do jerimum	1	6,53	1	F	paciente	3	6,45	1	F
chá de coentro	1	6,45	1	F	paciente diabético	2	6,45	1	F
coentro	2	6,45	1	V	papel	4	6,45	1	V
coisa	10	6,45	1	F	papel higiênico	1	6,45	1	F
colo	2	6,45	1	V	papelzinho de bombom	1	6,45	1	F
colo do útero	1	6,45	1	V	parceiro	3	6,45	1	F
colo uterino	1	6,45	1	V	pastoral da criança	1	6,45	1	F
comunidade	9	6,45	1	F	patologia	5	6,45	1	V
condições financeiras	1	6,45	1	F	pé da barriga	2	6,45	1	F
conhecimento	3	6,45	1	V	pedra	2	6,45	1	F
consciência	3	6,45	1	V	pedra dura	2	6,45	1	F
consciência da prevenção	1	6,45	1	F	pele	2	6,45	1	V
consciência da prevenção da gravidez	1	6,53	1	F	peso	6	6,45	1	F
controle de terrenos	1	6,45	1	F	peso das gestantes	2	6,45	1	F
controle do peso	1	6,45	1	F	pessoa	19	6,45	1	F
controle do peso das gestantes	1	6,53	1	F	pessoa de idade	1	6,45	1	F
coragem	2	6,45	1	F	pessoal	2	6,45	1	F
criança	19	6,45	1	V	pessoal da fns	1	6,45	1	F
cuidado	3	6,45	1	F	pessoal do eri	1	6,45	1	F
cuidado da criança	1	6,45	1	V	pivô principal	1	6,45	1	F
desnutrição	9	6,45	1	V	plantas medicinais	5	6,45	1	V

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
desnutrição da emoção	1	6,45	1	F	posto de saúde	5	6,45	1	F
desnutrição de alimentação	1	6,45	1	F	povo	2	6,45	1	F
desnutrição de alimento	1	6,45	1	F	pressão sangüínea	2	6,45	1	F
deus	2	6,45	1	F	prevenção	5	6,45	1	F
diabetes gestacional	1	6,45	1	V	prevenção da gravidez	1	6,45	1	F
diabético	5	6,45	1	F	prevenção de dst	1	6,45	1	F
diarréia	2	6,45	1	V	prevenção do câncer	2	6,45	1	F
dificuldade	2	6,45	1	F	profilaxia do diabético	1	6,45	1	F
dinheiro	3	6,45	1	F	profissional	4	6,45	1	F
doença	10	6,45	1	V	quadro	10	6,45	1	F
doença infecta	1	6,45	1	F	quadro clínico	5	6,45	1	F
dor	7	6,45	1	V	questionário	3	6,45	1	F
dor de cabeça	2	6,45	1	V	questionário de conduta	1	6,45	1	F
dor na mama	1	6,45	1	F	questionário de fitoterapia	1	6,45	1	F
dst	6	6,45	1	V	quilo	4	6,45	1	V
educação	3	6,45	1	V	quintal	4	6,45	1	F
efeito	2	6,45	1	F	rapaz	2	6,45	1	F
enfermeira encarregada	1	6,45	1	F	remédio	5	6,45	1	F
estudo do peso	1	6,45	1	F	riscos das doenças	1	6,45	1	F
estudo do peso das gestantes	1	6,53	1	F	rua	3	6,45	1	F
exame	10	6,45	1	F	rua familiar	1	6,45	1	F
exame ginecológico	1	6,45	1	F	saúde	27	6,45	1	V
exame laboratorial	1	6,45	1	F	secretaria de saúde	3	6,45	1	F
exames laboratoriais	2	6,45	1	F	secretaria de saúde do município	1	6,53	1	F
experiência	3	6,45	1	F	senhor	4	6,45	1	F
família	6	6,45	1	V	sexo	4	6,45	1	V
fato	3	6,45	1	F	soro caseiro	1	6,45	1	F
fator	4	6,45	1	F	terreno	3	6,45	1	F
fisioterapia	2	6,45	1	F	tratamento	4	6,45	1	F
fitoterapia	4	6,45	1	V	treinamento	5	6,45	1	F
flor do sabugueiro	1	6,45	1	F	uso	7	6,45	1	F
folha	6	6,45	1	F	uso da camisinha	1	6,45	1	F

Termo	Total	Tf-Idf	Entropy	Decs	Termo	Total	Tf-Idf	Entropy	Decs
folha da cenoura	2	6,45	1	F	vacina	2	6,45	1	F
folha do chuchu	1	6,45	1	F	vaso	2	6,45	1	F
folha do jerimum	1	6,45	1	F	vaso sanitário	2	6,45	1	F

Quadro 21: Resultado do experimento da área de inventario vocabular