



**Maria Cláudia de Freitas**

**Elaboração automática de ontologias de domínio:  
discussão e resultados**

PUC-Rio - Certificação Digital Nº 0310593/CA

**Tese de Doutorado**

Tese apresentada como requisito parcial para  
obtenção do título de Doutor pelo Programa de Pós-  
Graduação em Letras da PUC-Rio.

Orientador: Violeta de San Tiago Dantas Barbosa Quental

Rio de Janeiro, janeiro de 2007



**Maria Cláudia de Freitas**

## **Elaboração automática de ontologias de domínio: discussão e resultados**

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Letras da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

---

**Profa. Violeta de San Tiago Dantas Barbosa Quental**

Orientador  
Departamento de Letras – PUC-RIO

---

**Profa. Margarida Maria de Paula Basilio**

Departamento de Letras – PUC-RIO

---

**Profa. Helena Franco Martins**

Departamento de Letras – PUC-RIO

---

**Profa. Vera Lucia Strube de Lima**

Departamento de Fundamentos da Computação – PUC-RS

---

**Prof. Geraldo Bonorino Xexéo**

Instituto Alberto Luiz Coimbra de Pós-Graduação  
e Pesquisa de Engenharia – UFRJ

---

**Prof. Paulo Fernando Carneiro de Andrade**

Coordenador Setorial do Centro de Teologia e Ciências Humanas – PUC-RIO

Rio de Janeiro, \_\_\_\_\_ de \_\_\_\_\_ de \_\_\_\_\_.

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

### **Maria Cláudia de Freitas**

Graduou-se em letras (Português-Literatura) pela PUC-Rio em 1997. Obteve o título de Mestre em Letras pela PUC-Rio em 2000 e concluiu, em 2007, Doutorado em Letras (área de concentração: Estudos da Linguagem) na mesma instituição. Leciona na PUC-Rio desde 2002, ministrando cursos na área de Comunicação e Expressão, Lingüística e Língua Portuguesa. Participa, como pesquisadora, de projetos na área de Lingüística Computacional, desenvolvidos no CLIC - Centro de Lingüística Computacional da PUC-Rio.

#### Ficha Catalográfica

Freitas, Maria Cláudia de

Elaboração automática de ontologias de domínio :  
discussão e resultados / Maria Cláudia de Freitas ;  
orientadora: Violeta de San Tiago Dantas Barbosa Quental. –  
2007.

142 f. ; 30 cm

Dissertação (Mestrado em Letras)–Pontifícia  
Universidade Católica do Rio de Janeiro, Rio de Janeiro,  
2007.

Inclui bibliografia

1. Letras – Teses. 2. Ontologia. 3. Taxonomia. 4.  
Hierarquia lexical. 5. Extração de informação. 6. Relações  
semânticas. 7. Léxico. 8. Nomes próprios. I. Quental, Violeta  
de San Tiago Dantas Barbosa. II. Pontifícia Universidade  
Católica do Rio de Janeiro. Departamento de Letras. III.  
Título.

CDD: 400

## Agradecimentos

À Violeta Quental – pelo apoio, incentivo, amizade, generosidade, disponibilidade e, sobretudo, pelo bom humor e leveza com que trata o mundo acadêmico.

À Claudia Oliveira - pela generosidade, pela amizade, pelas discussões e por ter, em grande parte, viabilizado a interdisciplinaridade deste trabalho.

À Helena Martins – pela apresentação de “um outro ponto de vista” sobre a linguagem e pela preciosa – e luxuosa – assessoria teórica.

À Erica Rodrigues – pelo “SOS gramática” sempre disponível, pela amizade.

Ao Renato Paes Leme – pela paciência e prontidão com que transformava meus pedidos em um programa de computador.

Ao Cícero Nogueira dos Santos – pelo “suporte 24hs” nos sintagmas nominais, etiquetagens e afins, pelas dicas computacionais e pela paciência.

Ao Marcelo, à Ana e ao Raul – pelas muitas horas que passaram discutindo e avaliando listas de “X é um Y”.

Ao CLIC – pela troca valiosa de idéias.

À Chiquinha e à Dy – pela presteza, pela paciência e pelo sorriso.

## Resumo

Freitas, Maria Claudia de; Quental, Violeta de San Tiago Dantas Barbosa. **Elaboração automática de ontologias de domínio: discussão e resultados.** Rio de Janeiro, 2007. 142p. Tese de Doutorado - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

O objetivo deste trabalho é apresentar subsídios para a elaboração automática, a partir de corpus, de ontologias específicas quanto ao domínio. Para tanto, assumo que determinadas relações semânticas, como a hiperonímia, podem estar sistematicamente expressas em textos por meio de determinados padrões léxico-sintáticos. Tomando como ponto de partida alguns desses padrões, descritos originalmente em Hearst (1992, 1998), (i) identifico novos padrões para a expressão da relação de hiperonímia; (ii) adapto e refino três padrões já existentes (Hearst, 1992), tendo em vista especificidades da língua portuguesa; (iii) faço um cruzamento entre as informações extraídas com os padrões, a fim de gerar inferências. A perspectiva teórica subjacente é inspirada por reflexões wittgensteinianas sobre o significado, e se mostrou produtiva na medida em que legitima os dados vindos do corpus e as relações de significado que nele aparecem. O modelo de ontologia proposto caracteriza-se principalmente por: (i) não conter categorias pré-definidas, já que categorias são construtos humanos, abstrações que refletem uma perspectiva particular do mundo. A idéia de sustentar a ontologia em corpus busca deslocar o espaço de discussão sobre quais seriam as categorias relevantes de um domínio: as categorias que emergem do corpus refletiriam o conhecimento implícito do domínio em questão; (ii) não conter definições criadas a priori, sendo o significado de cada item decorrente das relações entre as palavras. A metodologia – extração das relações por meio de regras e posterior cruzamento para a realização de inferências – foi aplicada em um corpus do domínio saúde e um corpus genérico. Os resultados positivos indicam que sua utilização pode ser uma importante aliada na elaboração de ontologias e, também, uma ferramenta de auxílio a lexicógrafos e a sistemas de classificação semântica de nomes próprios. Em termos gerais, a metodologia apresenta como principais vantagens (i) a facilidade na automação do processo, minimizando a intervenção humana; (ii) facilidade na categorização de domínios

especializados; (iii) maior dinamicidade, pois o fato de o corpus poder ser constantemente atualizado faz com que esteja menos sujeito a falhas.

### **Palavras-chave**

ontologia; taxonomia; hierarquia lexical; extração de informação; relações semânticas; léxico; nomes próprios

## Abstract

Freitas, Maria Claudia de; Quental, Violeta de San Tiago Dantas Barbosa.  
**Elaboração automática de ontologias de domínio: discussão e resultados.**  
Rio de Janeiro, 2007. 142p. PhD - Departamento de Letras, Pontifícia  
Universidade Católica do Rio de Janeiro.

The main goal of this work is to present an automated method for building domain-specific corpus-based ontologies. The assumption is that semantic relationships, such as hypernym, can be systematically expressed through lexical-syntactic patterns. Starting with some of these patterns, originally described in Hearst (1992), I (i) identify new patterns that express hypernym; (ii) adapt three other patterns (Hearst, 1992), considering specificities of the Portuguese language; and (iii) intersect these results, in order to produce inferences. The theoretical approach is inspired by the wittgensteinian ideas about meaning. The resulting ontology's most prominent features are: (i) the fact that it does not have a priori categories, since categories are human constructs, abstractions that reflect a particular world view. Instead of discussing what should be the main categories in a domain, sustaining the ontology on corpora assumes that the corpus reflects the implicit knowledge of a given domain; and (ii) the fact that it does not have a priori definitions: the meaning of a word is derived from its relations with other words. The method – automatic extraction of semantic relations through rules, and the intersection of this information in order to produce inferences – was applied to two corpora: a health domain corpus and a generic corpus. The positive results show that the method can be very useful in ontology building and it can also be a valuable tool for lexicographers and named entity recognition systems. The main advantages of the method are (i) the simplicity of automating the process of ontology building; (ii) the ease of categorizing specialized domains, and (iii) its dynamicity, since the possibility of constantly updating the corpus makes it less subject to errors.

## Palavras-chave

ontology; taxonomy; lexical hierarchy; lexicon; proper nouns

## Sumário

1	Introdução	13
1.1.	Organização da tese	20
2	Um ponto de vista fértil	21
2.1.	O tratamento do significado no Processamento de Linguagem Natural	26
2.2.	Ontologias e significados – uma visão tradicional	30
2.3.	Ontologias e significado – uma visão relativista	33
2.4.	Ontologias, tesouros e taxonomias	35
2.5.	Sobre taxonomias e hipônimos	36
3	Crítérios para a elaboração e avaliação de ontologias	44
3.1.1.	Crítérios para a elaboração de ontologias “tradicionais”	44
3.1.2.	Crítérios para a elaboração de ontologias baseadas em corpus	46
3.2.	Formas de avaliação de ontologias	47
4	Trabalhos relacionados à extração automática de hiperonímia	54
4.1.	WordNet, EuroWordNet e Wordnet.Br	54
4.2.	Extração automática de hiperonímia	55
4.2.1.	Os padrões de Marti Hearst	56
4.2.2.	Outros trabalhos	60
5	Metodologia	66
5.1.	O corpus	66
5.1.1.	O pré-processamento do corpus	66
5.2.	Descrição dos padrões	67
5.2.1.	O padrão “tais como”	68
5.2.2.	O padrão “e/ou outros”	73

5.2.3. O padrão “tipos de”	75
5.2.4. O padrão “chamado/a/os/as”	76
5.2.5. O padrão “conhecido/a/os/as como”	76
6 Resultados	78
6.1. Análise dos erros sintáticos	79
6.2. Validação humana	83
6.2.1. Filtro 1: substantivos gerais	85
6.2.2. Filtros 2 e 3: adjetivos e pronomes	87
6.3. Novos resultados	90
6.4. Generalização e comparação dos resultados	92
7 Produzindo conhecimento novo: a realização de inferências	96
7.1. Inferências em um corpus genérico	106
7.2. Nomes Próprios	109
7.2.1. Classificação semântica de nomes próprios em um corpus genérico	112
8 Conclusões	116
8.1. Desdobramentos	121
8.1.1. Desdobramentos “mais” lingüísticos	121
8.1.2. Desdobramentos “mais” computacionais	122
8.2. Considerações finais	122
9 Referências bibliográficas	124
10 Anexos	130

## Lista de figuras

Figura 1: Categorias de Aristóteles, por Franz Bretano	31
Figura 2: Esquema conceitual como núcleo de um sistema integrado	32
Figura 3: Taxonomia de adoção produzida pela regra “hiperN”	97
Figura 4: Taxonomia de <i>áreas</i>	98
Figura 5: Taxonomia com inferência “artificial”	99
Figura 6: Taxonomia de <i>sintomas</i>	100
Figura 7: Diferentes contextos de uso de <i>drogas</i>	101
Figura 8: Taxonomia de <i>artrópodes</i>	102
Figura 9: Taxonomia de <i>conjunto</i>	102
Figura 10: Taxonomia de <i>estilos</i>	103
Figura 11: Recorte da taxonomia de <i>infecções</i>	103
Figura 12: Taxonomia de <i>objetos</i>	104
Figura 13: Taxonomia de <i>adornos</i>	108
Figura 14: Taxonomia de <i>estabelecimentos</i>	108
Figura 15: Taxonomia de <i>produtos</i>	108

## Lista de tabelas

Tabela 1: Resultados de busca na Internet por padrão discriminador	52
Tabela 2: Resultado da avaliação de 200 frases com o padrão “e outros” (Hearst, 1998)	59
Tabela 3: Resultados de alguns padrões de Morin e Jacquemin (2004)	61
Tabela 4: Resultados das extrações por padrão	79
Tabela 5: Análise dos erros sintáticos do padrão “como/tais como”	80
Tabela 6: Análise dos erros sintáticos do padrão “e/ou outros”	81
Tabela 7: Erros obtidos com o padrão “chamado”	82
Tabela 8: Resultados da avaliação humana	85
Tabela 9: Resultados da validação após aplicação dos filtros	90
Tabela 10: Resultados com o corpus genérico	92
Tabela 11: Comparação dos resultados	93
Tabela 12: Comparação entre os corpora com relação aos nomes próprios	112
Tabela 13: Resultados da avaliação de nomes próprios no corpus genérico	113

## Lista de quadros

Quadro 1: Lingüística baseada em corpus vs. Lingüística dirigida por corpus (Oliveira, 2006)	18
Quadro 2: Exemplos de etiquetas atribuídas ao “ <i>como</i> ” por etiquetadores automáticos	69
Quadro 3: Erros obtidos com o padrão “tipos de”	81
Quadro 4: Substantivos gerais eliminados	87
Quadro 5: Adjetivos mais freqüentes e de caráter geral	89
Quadro 6: Exemplos da aplicação do filtro de adjetivos	89
Quadro 7: Exemplos de relações que perderam especificidades com o filtro ADJ	90
Quadro 8: Processo de extração de relações de hiperonímia	91
Quadro 9: Resumo comparativo	95
Quadro 10: Taxonomias que produziram erros em decorrência de poslissemia	101
Quadro 11: Resultados da taxonomia no formato <i>bottom-up</i> para relações de 1 nível	105
Quadro 12: Resultados de visualização <i>bottom-up</i> para taxonomias com mais de um hiperônimo	106
Quadro 13: Visualização <i>top-down</i> de relações da amostra do CorpusCETENFolha	109
Quadro 14: relações extraídas de frases com ambigüidade no SPrep	113
Quadro 15: Relações corretamente extraídas que contêm SPrep.	113
Quadro 16: Resultados da categoria <i>empresas</i>	115
Quadro 17: Resultados da categoria <i>autores</i>	115
Quadro 18: Resultado da categoria <i>países</i>	115