Pontifícia Universidade Católica
do Rio de Janeiro

Maria Cláudia de Freitas

Elaboração automática de ontologias de domínio: discussão e resultados

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

Pontifícia Universidade Católica
do Rio de Janeiro

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 _____.

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

Lista de figuras

Lista de tabelas

Lista de quadros

1
Introdução

O objetivo deste trabalho é apresentar subsídios para a elaboração automática de ontologias específicas quanto ao domínio. Especificamente, busco investigar até que ponto é possível a elaboração automática de ontologias diretamente a partir de corpus, sem a determinação a priori das categorias que a compõem.

Vivemos na sociedade da informação, uma sociedade na qual o volume de informação nunca foi tão grande. Nossa capacidade de compreender, selecionar e organizar o conhecimento não consegue acompanhar a velocidade com que a informação, que aparece principalmente sob a forma de textos, é disponibilizada. Nesse contexto, é fundamental o desenvolvimento de ferramentas capazes de processar esse vasto material disponível; ferramentas capazes de extrair conhecimento de textos e transformá-lo em uma codificação da informação que seja armazenável, reutilizável e recuperável. E, mais ainda, de ferramentas voltadas para a língua portuguesa.

Objetivando auxiliar as tarefas de "gerenciamento" e "manipulação" da informação contida em textos, sistemas de recuperação e extração de informação têm se tornado populares. Porém, como afirma Vossen (2003), "para processar informação é preciso informação" (2003:464). Essa informação, por sua vez, pode ser mínima, vinda de um léxico que contenha apenas indicação sobre classes de palavras, ou pode ser de grande complexidade, quando originada de alguma base que contenha formalizações sobre conhecimento de mundo.

De fato, léxicos computacionais vêm assumindo crescente importância em sistemas de processamento automático de linguagem natural (PLN) (Boguraev e Pustejovsky, 1996). Um léxico computacional pode assumir tanto a estrutura linear de um dicionário quanto uma estrutura hierárquica, e, nesse caso, se aproximaria de uma taxonomia. Pode, também, fornecer outros tipos de relação, além da hiperonímia/hiponímia presentes em taxonomias. Quando o tipo de informação codificada é de natureza "mais lingüística", como a indicação sobre

classes de palavras, é comum o uso do termo léxico; para fazer referência a alguma base que contenha formalizações sobre conhecimento de (ou de algum) mundo, o termo ontologia costuma ser mais utilizado. Porém, como lembra Vossen (2003), a diferença entre léxicos e ontologias está longe de ser clara e há, sem dúvida, uma grande sobreposição sobre a informação que ambos veiculam.

Ao lidar com ontologias – descrições do mundo ou de porções do mundo – esta tese lida, indiretamente, com significado. Com isso, dialoga com a semântica, "domínio de investigação de limites movediços" e para o qual não há jargões bem estabelecidos (Ilari e Geraldi, 1985:6). Além disso, o trabalho se insere na área essencialmente interdisciplinar que é o Processamento automático de Linguagem Natural (PLN). Termos como ontologias, tesauros, léxicos semânticos e taxonomias são amplamente utilizados quando se quer fazer referência a bases que contêm informação sobre a língua necessárias ao processamento de textos, mas sua definição difere conforme o interesse e formação dos grupos de pesquisa, havendo pouca concordância sobre o que sejam.

Um léxico semântico, por exemplo, pode ser tanto uma lista de palavras com rótulos relativos à categoria semântica – a palavra *carro* pode ser rotulada como *veículo* (Phillips e Riloff, 2002; Riloff e Shepherd, 1997), quanto uma ferramenta responsável pela normalização entre termos e conceitos (Buitelaar, 2001). Isto é, assumindo que a língua é redundante, e que diferentes termos podem fazer referência ao mesmo objeto no mundo, a função de um léxico semântico seria realizar o mapeamento entre termos similares e conceito. Esta normalização deve considerar tanto informação relativa à classe semântica, definindo o tipo de objeto que um determinado termo ou conjunto de termos similares representa (por exemplo, *sinagoga*, *igreja* e *catedral* podem ser relacionadas à classe semântica *prédio religioso*); quanto informação relativa à estrutura semântica, definindo com quais outros objetos, atributos e ações tal objeto pode co-ocorrer (Buitelaar, 2001).

Dias-da-Silva (2004) apresenta seis definições para o termo tesauro:

- um inventário de itens do vocabulário de uma língua particular;
- um inventário de palavras tematicamente organizadas, isto é, um dicionário onomasiológico;
- um inventário de sinônimos e antônimos;
- um inventário que constitui um índice para a informação armazenada em um computador; uma lista de assuntos

relacionados à informação que deve ser recuperada por meio de palavras-chave;
— um inventário eletrônico, isto é, um arquivo de computador que armazena sinônimos que aparecem para o usuário durante o processo de correção automática;
— um inventário eletrônico de sinônimos e antônimos.

As *ontologias*, tema central desta tese, são de definição ainda mais variada. Como são objeto de estudo de diferentes áreas (filosofia, ciências cognitivas, inteligência artificial, semântica lexical, lexicografia e ciência da informação), é natural que haja uma multiplicidade de acepções que, não coincidentemente, corresponderão a diferentes tipos de ontologia. Brewster et al. (2005) chegam a afirmar que ontologias têm sido vendidas para a comunidade acadêmica como uma "panacéia" (2005:1). O termo pode fazer referência a taxonomias, como as do Yahoo, a bases de dados lexicais, como a WordNet (Fellbaum, 1998) e a construtos logicamente coerentes sobre os quais sistemas de raciocínio podem operar. Brewster e Wilks (2004) sugerem que tanto ontologias como taxonomias e tesauros estão dispostas em um continuum: em um extremo estariam as ontologias completamente explícitas, elaboradas de modo a facilitar o cálculo de inferências lógicas. Em outro extremo, estruturas que se organizam como mapas conceituais, que envolvem algum esforço de interpretação humana para que possam ser consideradas uma representação de conhecimento. Em algum ponto entre esses extremos estão taxonomias e hierarquias navegáveis na Internet, como os diretórios do Yahoo, claramente menos rigorosas do que uma ontologia completamente especificada. Os autores acreditam ainda que essas taxonomias "meio-termo" são, exatamente por não pretenderem total rigor teórico, mais fáceis de serem construídas de forma automática ou semi-automática.

Neste trabalho, utilizo a definição de Hovy (2002), segundo a qual uma ontologia é um conjunto de termos, associados a definições em linguagem natural, que utilizam, se possível, relações formais e restrições, sobre algum domínio de interesse, usado por humanos, bases de dados e programas de computador[1].

No âmbito do PLN, ontologias são úteis em uma série de tarefas. Na recuperação de informação e de documentos, ontologias permitem expansão do

[1] "For generality we define an ontology rather loosely as a set of terms, associated with definitions in natural language (say, English) and, if possible, using formal relations and

termo da busca, tanto por sinônimos quanto por hipônimos. Porém, é preciso considerar que este tipo de expansão, se, por um lado, leva a um aumento no número de documentos recuperados, por outro, leva a um declínio na precisão, isto é, mais documentos irrelevantes são recuperados.

Na sumarização automática, assume-se que frases que possuem palavras diferentes, mas relacionadas por meio de relações de hiperonímia ou sinonímia podem estar relacionadas, contribuindo para o cálculo de relevância de palavras em um texto. Ainda na área de geração de textos, a utilização de termos hiperônimos contribui para a coesão textual e maior fluidez do texto, evitando a repetição de palavras.

A resolução de anáforas é uma tarefa que também se beneficia de uma ontologia. Em um par de sentenças como "*Maria comprou pêssegos lindos. As frutas estavam doces e suculentas*", a relação de hiperonímia entre *pêssego* e *fruta* possibilita uma "compreensão" da sentença.

Atualmente, boa parte das aplicações de PLN que necessita de informação semântica utiliza a WordNet (Fellbaum, 1998). Porém, a WordNet é feita para a língua inglesa e, para o português brasileiro, embora o projeto Wordnet.Br (Dias-da-Silva, 2004) esteja em andamento, os resultados ainda não estão disponíveis para uso (a seção 4.1 trata detalhadamente das wordnets). Além da limitação imediata relativa ao idioma, outras restrições fazem com que o uso da WordNet como ontologia seja visto com ressalvas.

A primeira delas refere-se à presença freqüente de sentidos raros. A WordNet inclui, por exemplo, o sentido de *computador* como "*aquele que computa, que realiza cálculos*" e isso é um problema quando o objetivo é a expansão dos termos de uma busca, por exemplo, já que a expansão de *computador* incluirá sinônimos como *calculista* (Pantel e Ravichandran, 2004).

Outra limitação é a ausência de jargões, de termos específicos de determinadas áreas, bem como a presença esparsa de nomes próprios.

Por fim, e não menos importante: a WordNet é feita manualmente, o que implica um trabalho lento e dependente de vasta mão de obra. E, como o sucesso de um sistema é em grande parte dependente do tamanho da base, conseqüentemente é necessária uma grande equipe de pesquisadores e

---

constraints, about some domain of interest, used in their work by human, data bases, and computer

lexicógrafos para que ela seja efetivamente utilizada. Em conseqüência, sua atualização, um aspecto fundamental se admitimos que o conhecimento que se quer capturar está em constante fluxo, é mais custosa. Além disso, o caráter manual também esbarra nas limitações sofridas por dicionários: as definições estão sujeitas à subjetividade de lexicógrafos; ontologias e taxonomias refletem uma visão particular de mundo – a visão de quem as constrói, mesmo que corroborada por especialistas (Kilgarriff, 2003; Wilks, 2002).

Tendo em vista as restrições apresentadas, tem-se investido recentemente em formas de automatizar o processo de aquisição de informação lexical, desenvolvendo-se metodologias para a construção automática de bases de conhecimento, taxonomias ou ontologias (Hearst, 1992, 1998; Widdows, 2003; Snow et al., 2005; Phillips e Riloff, 2002; Caraballo, 1999; Maedche e Staab, 2000, entre outros), ancorando na informação contida em textos o conhecimento a ser representado (Buitelaar et al., 2005).

Uma ontologia como uma forma de representação do conhecimento é um modelo abstrato do que um indivíduo ou uma comunidade acreditam ser verdadeiro sobre o mundo. Nessa visão, textos seriam a única fonte concreta de informação com relação a esse conhecimento, na medida em que é possível sua análise, manipulação e extração de determinados tipos de informação (Brewster et al., 2005).

Dentre as propostas de construção automática de ontologias a partir de textos, não há investigações voltadas para a língua portuguesa. Este trabalho visa a suprir esta lacuna, apresentando subsídios para a construção automática de uma ontologia específica de domínio que auxilie o desempenho de tarefas de processamento automático de linguagem natural. Para tanto, proponho, seguindo os trabalhos para a língua inglesa desenvolvidos por Marti Hearst (1992, 1998), a extração de relações de hiperonímia em um corpus da área de saúde, por meio da identificação de determinados padrões léxico-sintáticos. Proponho, também, que os resultados obtidos nessa extração sejam cruzados de modo a possibilitar a realização de inferências – aumentando as informações disponíveis na ontologia.

Do ponto de vista teórico, assumo uma postura compatível com uma visão pragmática "radical" do significado, expressa sobretudo nas *Investigações*

---

programs" (Hovy 2002:92).

Filosóficas de Wittgenstein (1953), segundo a qual os significados não existem enquanto entidades autônomas.

Esta perspectiva também é compatível com a investigação do uso da língua em grandes corpora. A utilização de corpus para pesquisas lingüísticas pode ser compreendida tanto como uma metodologia quanto como uma teoria. Esta divisão encontra respaldo na distinção entre lingüística baseada em corpus e lingüística dirigida por corpus, notada em Sinclair (1996, apud Oliveira, 2006), como mostra o quadro 1, retirado de Oliveira (2006).

Lingüística baseada em corpus	Lingüística dirigida por corpus
o corpus é utilizado para validar, verificar e melhorar observações lingüísticas que já tenham sido realizadas	um corpus é de importância essencial no surgimento de novas idéias de como examinar os dados
o lingüista não questiona posições teóricas pré-estabelecidas ou categorias descritivas aceitas; sua posição com respeito à estrutura da língua já se estabilizou	o lingüista acredita que é possível conciliar o tipo de evidência que emerge do corpus com as posições estabelecidas; ele deixa abertas as possibilidades de mudanças radicais na teoria para lidar com as evidências
o corpus é utilizado para ajudar a estender e melhorar a descrição lingüística	a evidência do corpus é soberana, portanto o lingüista minimiza os pressupostos sobre a natureza das categorias teóricas e descritivas
um exemplo de questão relevante: WHOM ainda é utilizado em inglês? Como?	um exemplo de questão relevante: a distinção entre gramática e léxico é necessária?

Quadro 1: Lingüística baseada em corpus vs. Lingüística dirigida por corpus (Oliveira, 2006)

Segundo Oliveira (2006),

> "a distinção entre abordagens baseadas em corpus e dirigidas por corpus se assemelha ao contraste entre as abordagens top-down e bottom-up de resolução de problemas. No primeiro caso, o processo é analítico e os conceitos mais gerais da teoria do problema, suas abstrações de mais alto nível, são utilizadas para iniciar a análise. Os dados são os utilizados em última instância, na confirmação, extensão ou rejeição da teoria. Por outro lado, a abordagem bottom-up inicia-se com os dados e, em processos de síntese, formulam a teoria que abstrai e generaliza a informação inerente aos dados. Na prática da pesquisa lingüística, embora não na teoria, uma mistura das duas metodologias é invariavelmente necessária. No caso de uma pesquisa interdisciplinar, que busca meios lingüísticos de atingir objetivos computacionais, assim como prover meios computacionais para adicionar aos instrumentos de analise lingüística, a convergência das metodologias pode se acentuar"
>
> (Oliveira 2006:16)

Essa "mistura" das duas metodologias a que Oliveira se refere é o que Biber et al. (1998) chamam de abordagem baseada em corpus, uma abordagem que assume a complementaridade dos dois tipos de conhecimento.

Nesta tese, a abordagem baseada em corpus privilegia o trabalho de observação sobre o corpus na busca por determinados padrões léxico-sintáticos – isto é, privilegia o processo de síntese. Por outro lado, é inegável que o próprio *insight* sobre que tipo de padrão buscar, bem como sua formulação lingüística, só foram possíveis, ou melhor, foram bastante facilitados pela intuição da lingüista.

Em suma, a ontologia a ser desenvolvida apresenta as seguintes características:

- é baseada em língua;
- é totalmente baseada em corpus e não em dicionários ou outras bases preexistentes;
- é potencialmente infinita, pois novos termos podem ser constantemente acrescentados;
- é construída automaticamente.

A possibilidade de construção automática evidencia uma grande aproximação entre a metodologia proposta neste trabalho e técnicas da área de Extração de Informação.

A extração de informação (EI) pode ser considerada um tipo de recuperação de informação cujo objetivo é a retirada automática e seletiva de informações de documentos (textos). Trata-se de um processo que tem como entrada uma coleção de textos e que produz como saída dados em formato estruturado, que podem ser utilizados para povoar algum tipo de base de dados.

Vista desse modo, a tarefa de construção automática de ontologias pode ser considerada decorrência de técnicas de EI, pois se busca extrair, do texto, informação estruturada a respeito de determinadas relações entre as palavras. No caso específico dos padrões de hiperonímia apresentados aqui, uma grande vantagem é sua generalidade, que permite sua aplicação em diferentes domínios e gêneros textuais.

1.1. Organização da tese

No capítulo 2, trato dos fundamentos teóricos desta tese. Apresento o ponto de vista adotado para lidar com a questão dos significados e das relações semânticas entre as palavras. Ainda no capítulo 2, discuto as implicações das diferentes perspectivas sobre o significado para o entendimento do que são ontologias e analiso a visão tradicional a respeito de taxonomias e da relação de hiperonímia.

Os capítulos 3 e 4 constituem uma resenha da literatura sobre ontologias. No capítulo 3 apresento critérios formulados na tentativa de padronização do que sejam ontologias, dedicando atenção especial à proposta de Brewster e Wilks (2004), por tratar de ontologias construídas a partir de corpus. Além disso, examino as formas de avaliação que vêm sendo utilizadas na tentativa de aferição do sucesso e de comparação entre ontologias construídas automaticamente. No capítulo 4, descrevo os principais trabalhos que tratam da extração automática de relações de hiperonímia a partir de textos, apresentando de maneira mais detalhada as wordnets (ainda que estas não sejam feitas automaticamente) e o trabalho de Marti Hearst (1992, 1998).

Os capítulos 5 e 6 são o cerne deste trabalho. No capítulo 5, descrevo a metodologia: o corpus e os padrões utilizados na identificação da hiperonímia; e no capítulo 6 apresento os resultados obtidos.

Por fim, no capítulo 7, reflito sobre a proposta inicial – a possibilidade de elaboração automática de ontologia específica de domínio a partir de corpus – à luz dos resultados obtidos e apresento sugestões de trabalhos futuros.

2
Um ponto de vista fértil

Esta tese trata da elaboração automática de ontologias, inserindo-se na linha de pesquisa de processamento de linguagem natural (PLN). Uma vez que ontologias dizem respeito à descrição do mundo (ou de porções dele), e que "o projeto de dizer o que uma coisa é coincide inescapavelmente com a tentativa de circunscrever o significado de uma expressão lingüística" (Martins, 1999:137), a tese trata também, ainda que tangencialmente, de questões relacionadas ao significado, aproximando-se então do terreno movediço da semântica.

A questão "o que é o significado de uma palavra" é um dos problemas nucleares da investigação semântica. De forma bastante simplificada, é possível distinguir três paradigmas que irão problematizar o significado de forma sistemática: realista, mentalista e pragmática. Porém, ainda que didaticamente esta distinção seja útil, na prática, teorias realistas e mentalistas têm historicamente compartilhado pressupostos teóricos fundamentais, o que permite, com alguma simplificação, agrupá-las sob o rótulo *representacionistas* ou *essencialistas* (Martins, 2004).

Em uma visão mentalista, as palavras possuem uma relação estável com entidades mentais, isto é, a um significado corresponde um conceito, uma idéia. Já em uma visão realista, as palavras possuem uma relação estável com a realidade, com entidades do mundo que, por sua vez, podem ser reais ou virtuais. Para ambos, a linguagem é um sistema de representações de significados fixos e compartilhados; palavras representam *algo* (entidades mentais para os primeiros e virtuais para os segundos), e essa relação de representação se dá de maneira objetiva e estável.

Já o ponto de vista pragmático diz respeito à linguagem em uso, em diferentes contextos, considerando o uso feito pelos falantes na comunicação – o foco está na linguagem enquanto forma de *interação social*. O significado é decorrência de situações concretas, variáveis. Há, portanto, uma mudança de perspectiva, já que a linguagem passa a ser entendida como uma prática

intersubjetiva. Dentre as linhas de investigação pragmáticas, contudo, há as que poderiam ser também enquadradas em um paradigma essencialista. Isto porque se, por um lado, mentalistas irão assumir que é pela análise das propriedades dos códigos de linguagem que será possível explicar a prática da comunicação, algumas correntes da pragmática recomendam a análise das propriedades da prática da comunicação como maneira de fornecer uma explicação do que são as línguas e os significados, o que faz com que esta visão pragmática tradicional possa ser compreendida como uma disciplina complementar a uma visão semântica essencialista (Martins, 1999; Taylor, 1992). Porém, a crítica que pragmatistas mais radicais farão é que qualquer análise essencialista da linguagem é impossível, por ser impossível um distanciamento do objeto examinado; há uma relação mútua indissociável – nossas práticas de vida constituem a linguagem e, ao mesmo tempo, são por ela constituídas, o que impossibilita a realização de julgamentos absolutos sobre a linguagem. A relação entre linguagem e realidade seria forjada, na medida em que a própria linguagem constitui a realidade:

> "o que está sobrando é a pergunta 'Como a linguagem se liga à realidade?'. Pois se baseia firmemente em uma linguagem equivocada."
>
> (Hacker e Backer, 1984a :135)

Assumo neste trabalho uma postura compatível com uma visão pragmática radical do significado, segundo a qual a dificuldade em se responder à pergunta *o que é o significado* se deve à natureza equivocada da pergunta. A linguagem diz não o real em si, mas as opiniões e práticas dos homens, e por isso sua imprevisibilidade não é um desvio, mas conseqüência dessas opiniões ou impressões, que são naturalmente contraditórias (Martins, 2004).

Para lidar com o significado no ambiente do PLN, me apóio principalmente no ângulo sugerido por Wittgenstein, sobretudo nas *Investigações Filosóficas* (1953). É importante salientar, contudo, que Wittgenstein não apresenta uma teoria semântica, uma teoria do significado, ou mesmo uma filosofia da linguagem. Uma de suas grandes preocupações é mostrar que a linguagem não é um fenômeno único, e se furta a generalizações e sistematizações; o que ele propõe é uma elucidação do significado das palavras por meio da descrição de seu uso. Ao apontar para a resistência da linguagem à

investigação científica, Wittgenstein parece tematizar especificamente a questão do sentido na linguagem, sugerindo a inadequação da busca por uma ciência do significado (Martins, 1999).

Porém, assumir a inadequação da questão *o que é significado* não significa a defesa de uma posição reducionista segundo a qual significados não existem. Eles existem, mas não como entidades autônomas, e não com a precisão ou os limites definidos, necessários à formalização que sempre se buscou fazer. O significado é flexível e maleável, não cabe no molde fixo que lhe desejam impor. E, se esta recusa dos significados a uma formalização exaustiva pode ser uma forte limitação para as semânticas formais, por outro lado, pode representar uma motivação para outras formas de lidar com o significado. O significado não é uma propriedade imanente à palavra, mas uma função que expressões lingüísticas exercem em um contexto específico e com objetivos específicos (Marcondes, 2005). Com isso, o significado de uma palavra pode variar conforme o contexto em que é utilizado, conforme o objetivo desse uso.

Se não há uma essência única e fixa do significado, como lidar com as definições? Dicionários não só existem como são úteis. Negar esse fato parece um contra-senso. Porém, o que Wittgenstein enfatiza é o caráter parcial e incompleto das definições – que nem por isso as torna menos úteis. Desse modo, se, em uma perspectiva essencialista, esbarraríamos, em algum momento, nos "indefiníveis" – isto é, traços ou universais como "humano" ou "masculino", que compreendemos sem dificuldade – Wittgenstein argumenta que as definições são sempre fundamentadas em um conhecimento prévio, derivado do uso (do contexto, da situação de explicação, de inúmeros outros fatores). Isto é, definições, embora úteis nos contextos em que são utilizadas, serão sempre parciais. Explicações são sempre correlatas a pedidos de explicação, de modo que o significado é explicitado principalmente em situações que buscam desfazer mal-entendidos:

> Isso será feito (a descrição do uso de uma palavra, dizendo que objeto essa palavra designa) quando se tratar apenas de desfazer o mal entendido seguinte: pensar que a palavra lajota se relacione com a forma da pedra de construção que nós de fato nomeamos "cubos"' – mas o modo dessa 'relação', isto é, o uso dessas palavras, no restante, é conhecido."
>
> (Investigações Filosóficas, § 10)

Definições analíticas, que analisam termos com base em uma conjunção de marcas características, deixam de ser encaradas como "as" definições por excelência: trata-se apenas de mais uma forma de explicação, dentre outras possíveis. E, justamente por ser dependente do uso, dependente de uma situação concreta, e não uma entidade autônoma, a descrição do significado de um termo dificilmente se adequará ao formato das definições analíticas, composicionais e essenciais. Tais estratégias serão sempre limitadas:

> E o que ocorre com a última elucidação dessa cadeia? (Não diga "Não há nenhuma 'última' elucidação". É exatamente o mesmo que dizer: "Não há nenhuma última casa nesta rua ; pode-se sempre construir mais uma".)
>
> (Investigações Filosóficas, § 29)

É importante frisar que Wittgenstein não nega a validade de definições analíticas – definições analíticas são apenas um dos tipos possíveis de explicação, e enquanto tais são lances legítimos no *jogo de linguagem*[2] - , apenas lembra que elas são parciais, e não fundacionais na linguagem. Isto porque é impossível tomar distanciamento no jogo, isto é, parar de jogar e observá-lo de um ponto de vista exterior. Podemos fornecer explicações, generalizações, mas tudo isso consiste, ou já está previsto, no próprio jogo. Explicações, portanto, enquanto lances no jogo, funcionarão, isto é, servirão aos objetivos pretendidos, quando aplicadas às situações em que são produzidas, e não em todas as situações possíveis. Por isso, não são exaustivas, não são completas em si mesmas, não são absolutas (Martins, 2000).

Nesse sentido, a incompletude inerente às definições é uma faceta da ausência de um ideal de exatidão. Precisão e exatidão, novamente, são relativos. Não há um padrão único que as governe; a precisão é uma questão de adequação às circunstâncias e aos propósitos.

---

[2] O termo *jogo de linguagem* é utilizado de diferentes maneiras, em diferentes situações, sem, contudo, jamais ser explicitamente definido. Como observa Perloff, o termo é "difícil de entender, de vez que Wittgenstein, como é típico, jamais o define de forma plena, optando, em vez disso, por usá-lo freqüentemente, de um modo que ele acaba por tornar-se nosso" (Perloff, 1996:60, apud Martins, 1999:154). Detenho-me aqui no uso da expressão enquanto forma de "enfocar mais de perto as nossas atividades lingüísticas reais, descrevendo-as contra o pano de fundo de nossas práticas não lingüísticas" (Glock, 1997: 228). Fazem parte dos jogos de linguagem *atos de fala*; atividades mais complexas como *contar histórias*, *formar hipóteses e*

> "É inexato se eu não indicar a distância que nos separa até o sol até exatamente 1 m? E se eu não indicar ao marceneiro a largura da mesa até 0,001 mm?
> **Um** ideal de exatidão não está previsto; não sabemos o que devemos nos representar com isso – a menos que você mesmo estabeleça o que deve ser assim chamado. Mas ser lhe á difícil encontrar tal determinação; uma que o satisfaça."
> (Investigações Filosóficas, § 88)

A língua é naturalmente vaga, imprevisível e ambígua, e grande parte de sua robustez se deve justamente a isso. Nem todos os conceitos, porém, são realmente vagos, e, embora a maior parte dos conceitos empíricos admita casos fronteiriços, nem por isso se tornam inúteis (Glock, 1997).

> "It is precisely the lack of clarity in our use of the word culture which makes it such a handy word to have at one's disposal."
> (Stock, 1983 apud Kilgarriff, 1997: 39)

Na transposição das idéias de Wittgenstein para a lingüística, sigo aqui o caminho apresentado por Helena Martins (1999), segundo o qual uma lingüística descritiva compatível com o espírito wittgensteiniano ambiciona

> "fornecer descrições parciais e contingentes de regularidades observáveis nas línguas do mundo – sem pretender dar conta dos jogos como um todo, a partir de algum ponto de vista exterior(...)"
> (Martins, 1999:144)

Uma lingüística que

> "ambicionará, em seu impulso genuína e legitimamente generalizador, manter-se nos limites da linguagem, apresentando não uma visão verdadeira e completa dos fatos, antes um ângulo fértil pelo qual se possam reconhecer regularidades em nossas práticas lingüísticas."
> (Martins, 1999:144)

A perspectiva de Wittgenstein, por assumir nossa imersão no jogo da linguagem, é capaz de acomodar um ecletismo, uma visão mais tolerante com relação às diferentes teorias de linguagem. Com isso possibilita o uso, por exemplo, de um vocabulário tradicional, compreendido como lance no jogo de linguagem – do jogo de falar sobre a linguagem. Conseqüentemente, embora adotando o ponto de vista wittgensteiniano, não me privo, em diversos momentos,

---

testá-las; modos de discurso como *falar sobre objetos físicos* e jogos de linguagem de *falar sobre a linguagem* (Glock, 1997; Martins 1999).

da utilização de um vocabulário tradicional – em especial, de termos como *sintagma nominal*, *palavras denotativas*, *vagueza* e *hiponímia* – embora estes devam ser compreendidos de maneira deflacionada. Como bem esclarece Martins:

> "qualquer teoria sobre as línguas naturais será uma descrição parcial e reificadora de práticas sociais humanas – e isso vale tanto para as produzidas segundo um ângulo estruturalista quanto para aquelas que adotam o ponto de vista contextualista. (...) Continua fazendo algum sentido dizer, afinal, que o sistema verbal do português divide-se em três conjugações, ou que algumas línguas favorecem a omissão de sujeito na frase (...). Sem explicitar de maneira direta relações entre o lingüístico e o contextual, essas generalizações obviamente lançam alguma luz sobre nossos jogos de linguagem; se não alcançam a meta de revelar regras apriorísticas definitivas, pelo menos constituem descrições de regularidades envolvidas em nossas práticas lingüísticas."
>
> (Martins, 1999:146-147)

Lembro, por fim, que a idéia de que *explicações são sempre correlatas a pedidos de explicação* deve ser entendida de maneira abrangente. Assim, assumo aqui que o meu "*pedido de explicação*" é uma *aplicação* – uma ontologia específica de domínio. As explicações oferecidas, portanto, não pretendem um esgotamento da questão, mas pretendem responder, de maneira satisfatória, ao pedido.

As relações capturadas entre as palavras[3] retratam descrições parciais e contingentes de modos como são usadas nos jogos de linguagem em uma determinada língua (Martins, 1999), e acrescento, no caso específico deste trabalho, em um determinado domínio. Categorizar também é um jogar um jogo de linguagem com palavras.

2.1. O tratamento do significado no Processamento de Linguagem Natural

Tradicionalmente, a semântica computacional está ancorada em visões essencialistas-representacionistas do significado. As wordnets (Fellbaum, 1998; Vossen, 1998), por exemplo, são bases de dados lexicais que contêm "nomes,

[3] A delimitação de unidades lexicais é um tema controverso na teoria lingüística (Basílio, 1999). Neste trabalho, uso indistintamente "palavra" e "termo" para fazer referência às unidades lexicais.

verbos, adjetivos e advérbios agrupados em conjuntos de sinônimos cognitivos, *cada um representando um conceito distinto*"[4] (grifo meu).

Porém, nem sempre a diferença de abordagens com relação ao significado é nítida na Inteligência Artificial. Modelos como redes semânticas e enquadres fazem uso de inserção de conhecimento enciclopédico por um lado – incorporando elementos da pragmática tradicional – , e do formalismo de definição por traços e primitivos semânticos, por outro – incorporando elementos de um paradigma representacionista.

Além disso, se, como afirma Martins, "as idéias de Wittgenstein não têm comparecido em teorias lingüísticas com muita freqüência" (1999:136), no PLN a situação não é muito diferente – o que de forma alguma chega a ser surpreendente, visto a posição de Wittgenstein de não oferecer qualquer teoria unificadora sobre a linguagem e, principalmente, sua concepção de linguagem enquanto prática de vida que dificulta, por motivos óbvios, a possibilidade de transposições bem-sucedidas.

Em geral, o ponto de vista wittgensteiniano irá influenciar abordagens estatísticas do significado, especificamente aquelas voltadas para as tarefas de desambigüização de itens lexicais, em grande parte devido ao slogan "o significado está no uso". Nestes casos, a aproximação se dá por meio da substituição de *uso* pelo corpus; especificamente, pelas adjacências de uma palavra. Em termos gerais, calcula-se, para uma dada palavra-alvo, o número de palavras que aparecem ao seu lado em uma janela de tamanho pré-determinado - por exemplo, 15 palavras. Em seguida, cada palavra é representada por meio das freqüências cumulativas das ocorrências no escopo da janela. Palavras com significados similares tenderão a ocorrer em contextos similares e palavras polissêmicas tenderão a ocorrer em contextos diferentes.

Porém, as aproximações entre este tipo de trabalho e uma posição relativista com relação ao significado devem ser vistas com alguma cautela. Freqüentemente, o slogan serve de fachada para um trabalho estatístico que opta pela praticidade da não-definição dos termos. Schütze (1998), que propõe um mecanismo automático para a discriminação de significados, e não para a desambiguação, justifica sua

[4] Disponível em http://wordnet.princeton.edu/. Acessado em 19/12/2006

escolha exatamente por ser a desambiguação dependente de uma definição do significado[5].

Widdows (2003) apresenta um modelo de aquisição e desambigüização lexical baseado em informação estatística contextual, no qual não existem definições de palavras, mas apenas relações entre os termos. Mas, embora dispense as definições, ele afirma que o significado deve poder ser descrito de forma "clara, flexível e acurada", através de um pensamento científico cuidadoso e de investigação empírica. Ainda segundo Widdows (2003), métodos estatísticos, embora tenham trazido enormes contribuições, apenas *imaginam* ou *supõem* o significado das palavras.

Os trabalhos de Adam Kilgarriff, claramente inspirados em posições relativistas do significado, são os mais afinados com a perspectiva teórica assumida aqui. No artigo "I dont believe in word senses" (1997), Kilgarriff, tomando o ponto de vista de um lexicógrafo, salienta que significados só existem com relação a um objetivo, que pode ser o de escrever dicionários ou tesauros, por exemplo. Lembrando ainda a escassa literatura sobre a utilização de critérios na tarefa subjetiva de separação dos significados das palavras – o que contribui para dificultar ainda mais o trabalho dos lexicógrafos –, Kilgarriff sustenta que uma lexicografia de corpus é a mais apropriada para o tratamento dos significados, uma vez que ofereceria uma resposta diferente para a questão do significado de uma palavra. Assumindo, com Wittgenstein, que as palavras só têm sentido no uso, o lexicógrafo deve recorrer ao corpus como se fosse ele, o lexicógrafo, um "instrumento" cuja função é organizar o que está no corpus e "traduzir" esta organização para a linguagem de definição de dicionário.

> "Word senses are simply undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not worth making"
>
> (Kilgarriff, 1997: 107)

Em outro trabalho, voltando-se diretamente para o PLN (2003), Kilgarriff propõe que tesauros sejam construídos automaticamente a partir de corpora, com base não nos diferentes significados das palavras, mas nas próprias palavras. Por

[5] "Word sense discrimination is easier than full disambiguation since we need only determine which occurrences have the same meaning, and not what the meaning actually is"

meio da aplicação de algoritmos de agrupamento (*clustering*) sobre um corpus, o tesauro seria um agrupamento de termos relacionados – e o significado seria atribuído à palavra em função do grupo a que a palavra pertence.

Os trabalhos de Yorick Wilks também buscam aproximações com uma visão não-representacionista (Wilks, 1999; Niremburg e Wilks, 2001), e ultimamente Wilks tem se dedicado à investigação de ontologias (Wilks, 2002; Niremburg e Wilks, 2001; Brewster e Wilks, 2004).

Com relação à língua portuguesa, os trabalhos de Garrão (Garrão et al., 2006; Garrão, 2006), voltados para a identificação de expressões multi-vocabulares verbais, também incorporam um ponto de vista não-representacionista semelhante ao apresentado aqui.

Por fim, embora a enorme afinidade entre a perspectiva que assumo neste trabalho – de natureza predominantemente aplicada – e a crença de que esta perspectiva, principalmente nos termos de Martins (1999) é, de fato, promissora para estudos relativos ao significado, não acredito que esta afinidade seja condição necessária para o sucesso da investigação em PLN. Neste ponto, compreendo que um dos objetivos do PLN é a resolução de problemas. Assumo aqui, portanto, a perspectiva da IA fraca: basta que o desempenho dos programas imite o funcionamento da linguagem, não é preciso que os processos subjacentes, em ambos os casos, sejam os mesmos. Meu comprometimento, nesse sentido, é com resultados satisfatórios, e não com determinadas perspectivas teóricas, as quais são utilizadas na medida em que oferecem *insights* interessantes para a abordagem dos problemas. Concordo, portanto, com Diana Santos quando afirma que

> "(...) é ao tentar resolver um dado problema (isto é, ao tentar construir um programa que manipula a língua) que surge o momento de nos debruçarmos quer sobre (algumas características) do léxico ou da gramática, quer sobre as teorias que pretendam dar respostas a esse problema"
>
> (Santos, 2001:229)

A perspectiva wittgensteiniana de linguagem é compatível com essa visão, já que, para Wittgenstein, não é possível uma descrição completa da língua porque não é possível deixarmos de tomar parte no jogo para apenas observar; não é

(Schütze, 1998:97)

possível termos a visão do todo. Conseqüentemente, não há objeto ou processo a ser simulado. De fato, como afirma Sparck Jones,

> "The challenge of taking the necessary step from a focused experiment or even convincing prototype to a full-scale rounded-out NLP system with consistent, high-quality performance has not been overcome."
>
> (Sparck Jones, 2001:9)

Em conseqüência, acredito que a perspectiva teórica adotada para cada situação problema indica apenas que ela foi a mais produtiva para o tratamento daquele problema específico, mas não necessariamente que o será em outros casos. Enfim, para o tratamento de relações de significado entre as palavras, uma visão não-representacionista é um ponto de vista fértil.

2.2. Ontologias e significados – uma visão tradicional

O estudo das ontologias, embora desperte grande interesse no campo da Inteligência Artificial (IA), remonta às origens da filosofia, há cerca de 25 séculos. Mas esta longa tradição não significa que existam respostas satisfatórias aos problemas inicialmente apresentados. O termo, originalmente, designa o estudo do ser, considerado independentemente de suas determinações particulares e naquilo que constitui sua inteligibilidade própria. Trata-se da teoria do ser em geral, da essência do real (Japiassú e Marcondes, 1989). Enquanto teoria do ser, uma ontologia busca descrever as categorias mais básicas da realidade - entidades, tipos de entidades e o relacionamento entre esses elementos.

A investigação sobre as categorias que compõem a realidade começa a receber um tratamento sistematizado nas *Categorias*, de Aristóteles, que apresenta 10 categorias básicas que classificariam tudo o que pode ser dito ou predicado sobre qualquer coisa: substância, quantidade, qualidade, relação, lugar, tempo, posição, estado, atividade e passividade. O filósofo Franz Bretano, em 1862, adicionou alguns termos retirados de outros escritos de Aristóteles, e criou um diagrama de árvore como o da figura 1 (apud Sowa, 1999).

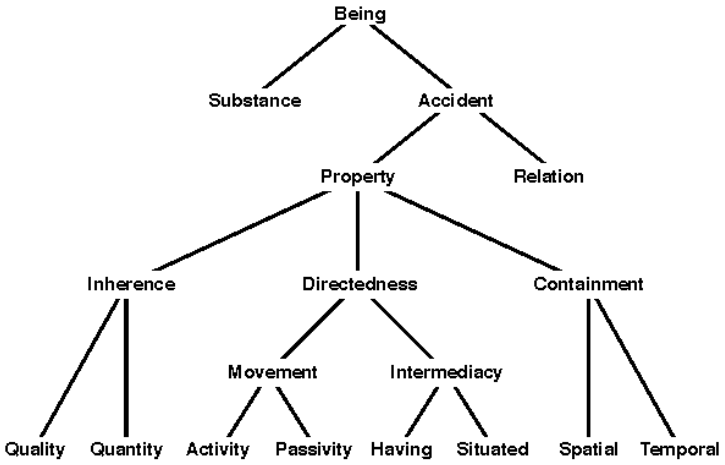


Figura 1: Categorias de Aristóteles, por Franz Bretano (apud Sowa, 1999)

As categorias expressas pela realidade descreveriam o real – assume-se a existência de um mundo externo à linguagem, passível de descrição. Ontologias devem, portanto, ser gerais e independentes de língua, pois descrevem a realidade, que, por sua vez, é a mesma para todos – e por isso os conceitos são gerais, independentes de língua. Ou seja, nessa visão, aos conceitos das ontologias são atribuídos rótulos – as palavras – , que serão dependentes de língua. De fato, essa é a perspectiva que norteia, até hoje, redes lexicais como as wordnets (Fellbaum, 1998; Vossen, 1998), que freqüentemente são utilizadas como ontologias:

> "In principle, the separation between ontology and lexicon is as follows: 'language-neutral' meanings are stored in the former; language-specific information in the latter."
>
> (Viegas et al., 1999: 21)

Na IA, a necessidade de formas padronizadas para a codificação do conhecimento foi reconhecida no início da década de 70. O ANSI (*American National Standards Institute*) propôs que todo o conhecimento pertinente sobre um domínio deveria estar concentrado em um único *esquema conceitual*, como ilustra a figura 2 (apud Sowa, 1999). A função de tal esquema seria fornecer definições comuns para as entidades das aplicações e explicitar o relacionamento entre elas (Sowa, 1999).

De acordo com Sowa (1999), por mais de 20 anos esse esquema conceitual foi importante no desenvolvimento e uso de aplicações integradas, mas nunca

houve implementações completas; nunca se atingiu o objetivo final de integração total em torno de um único esquema.

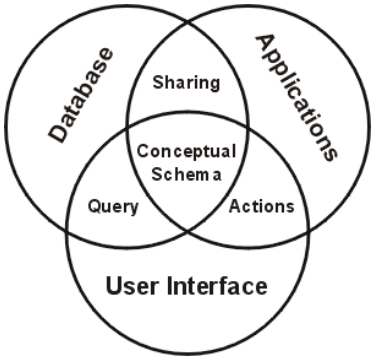


Figura 2: Esquema conceitual como núcleo de um sistema integrado (Sowa, 1999)

É nesse contexto que a IA se apropria do termo ontologia: o crescente reconhecimento de que fontes computacionais devem ser as mais gerais possíveis, reutilizáveis e compartilháveis entre a comunidade de IA foi o primeiro passo para considerar o valor das questões tradicionais da filosofia: o estudo da realidade e seus objetos, independentemente do nosso conhecimento sobre eles, e a busca por uma natureza a priori das coisas (Bateman, 1995). Para Guarino (1995), uma base de conhecimento que se aproximasse à noção filosófica clássica de verdade facilitaria não apenas a interação e comunicação entre diferentes agentes, mas também o compartilhamento e reaproveitamento da própria base.

Segundo Bateman (1995), no que tange às ontologias, há uma confluência apenas aparente de interesses entre filosofia e IA. Na IA, o uso do termo remeteria à construção de *frameworks* para "conhecimento" que permitam a sistemas computacionais lidar com problemas tais como processamento de linguagem natural e "*real world reasoning*". De acordo com essa perspectiva, um sistema deve ser capaz de realizar deduções com relação a algum corpo de informação, e os componentes organizacionais mais gerais desta informação são chamados coletivamente de ontologia. Guarino (1995) defende a introdução sistemática de princípios de ontologia formal na engenharia de conhecimento, a fim de explorar as várias relações entre ontologia e representação de conhecimento. Para a área de sistemas de informação, uma ontologia seria uma linguagem formal elaborada para representar um domínio particular de conhecimento, cujo objetivo é, essencialmente, funcional (Zúñiga, 2001). Em última análise, a própria discussão

sobre o que venha a ser uma ontologia é ilustrativa da dificuldade de se estabelecerem definições e conceitos comuns e compartilháveis entre domínios. Ou seja, a dificuldade em se chegar a um acordo sobre o que são ontologias põe em xeque a própria existência de ontologias nos moldes propostos – uma ontologia geral, multilingüe e, algumas vezes, independente de domínio.

De fato, a elaboração de ontologias sustentadas por representações de conhecimento gerais, independentes de língua, parece ser problemática. O projeto de construção de uma única ontologia, que pudesse ser ao mesmo tempo não-trivial e adaptável para diferentes comunidades de sistemas de informação, foi em grande parte abandonado; a tarefa se mostrou muito mais difícil do que o previsto inicialmente, confirmando os problemas já enfrentados por filósofos há 2000 anos (Smith, 2001).

O desapontamento com construção de ontologias gerais, levou, por sua vez, ao investimento em ontologias específicas de um domínio. Neste contexto, uma das definições de ontologia mais difundidas é a de Gruber (1993), segundo a qual uma ontologia é "uma especificação formal explícita de uma conceitualização compartilhada".

No âmbito da pesquisa em PLN, ontologias podem ser vistas como "modelos de domínios específicos", que têm como objetivo facilitar buscas semânticas (Brewster e Wilks, 2004).

2.3. Ontologias e significado – uma visão relativista[6]

Paralelamente à visão tradicional, desenvolve-se, na filosofia, uma outra abordagem, relativista, anti-essencialista, cujo embrião pode ser encontrado já no pensamento sofista, e que sustenta não existir uma realidade independente e exterior à linguagem e, portanto, passível de uma descrição essencialista. Segundo essa perspectiva pragmática radical, a própria empreitada ontológica perde o sentido – isto é, não se trata de uma tarefa difícil, mas de uma tarefa sem sentido: não há conceitos independentes de língua que descrevem o universo (ou parte dele) – em última análise, não há universo a ser descrito independente de língua, vista como práxis. O estabelecimento de verdades universalmente válidas,

[6] Refiro-me, no decorrer do trabalho, a um relativismo lingüístico-conceitual.

autônomas com relação às circunstâncias concretas é, do ponto de vista wittgensteiniano assumido neste trabalho, impossível. Somos constituídos pela linguagem, o que impossibilita a realização de julgamentos absolutos sobre ela. Ontologias gerais, aproximações às noções de verdade, não são questões que devam ser consideradas.

Mas, se não há "entidades mentais" ou realidade às quais as palavras se "colam", e que corresponderiam ao significado das palavras, o que é o significado então? A posição anti-essencialista de Wittgenstein, expressa principalmente nas *Investigações Filosóficas* (1953) e abordada no início deste capítulo, é de grande valia para lidar com o significado – intimamente relacionado à questão da elaboração de ontologias. Os significados correspondem aos usos culturalmente determinados que fazemos das palavras – o significado não é uma entidade, ele está no uso (Martins, 2004).

E no que as considerações de Wittgenstein podem ser úteis à semântica computacional, à elaboração de ontologias?

Na IA, como já mencionado, a ambição inicial de ontologias gerais foi substituída pela idéia de ontologias de domínio. Além da redução no escopo da tarefa, a constatação de que a elaboração de ontologias exige um processo longo de concordâncias entre um número grande de especialistas levou à pesquisa sobre formas de automação desse processo, considerando-se que o conhecimento a ser representado na ontologia deve ser a informação contida em textos (Buitelaar et al., 2005).

Adotando uma perspectiva relativista, na qual a linguagem e realidade se constituem mutuamente, é difícil pensar em ontologias baseadas em conceitos pré-definidos. Por outro lado, é igualmente difícil transpor a "linguagem enquanto prática de vida" para um ambiente computacional. Diante desse impasse, proponho a substituição (grosseira, é verdade) de "práticas de vida" pela informação contida no corpus – assumo que o conhecimento disponível em textos, expresso em linguagem natural, pode funcionar como uma fonte confiável para a busca de informações e categorizações.

Conseqüentemente, a principal característica da ontologia proposta é a ausência de categorias pré-definidas. Categorias em uma taxonomia são construtos humanos, abstrações que refletem uma perspectiva particular do mundo (Kilgarriff 2003, 1997; Brewster e Wilks, 2004). A idéia de sustentar a ontologia

em corpus busca deslocar o espaço de discussão sobre quais seriam as categorias relevantes de um domínio: do desejado consenso entre especialistas para as categorias motivadas pelo corpus, que, por sua vez, refletiriam o conhecimento implícito do domínio em questão.

2.4. Ontologias, tesauros e taxonomias

Como dito anteriormente, ontologias, no contexto de PLN, podem ser vistas como "modelos de domínios específicos", que têm como objetivo facilitar buscas semânticas. Neste ponto, surge outra confusão terminológica: no que diferem ontologias, tesauros, taxonomias e hierarquias? Trata-se de objetos cujas características se sobrepõem e que também compartilham, no PLN, do mesmo objetivo: auxiliar buscas semânticas. Em conseqüência, encontramos trabalhos muito parecidos mas que atribuem diferentes nomes aos seus "modelos": ora fala-se em tesauro (Kilgarriff, 2003), ora em ontologia (Vossen, 2003; Velardi et al., 2005; Brewster et al., 2005) e ora em taxonomia (Snow et al., 2005, Widdows, 2003).

Depois da discussão sobre o uso do termo ontologia apresentada na seção 2.2, volto ao tema, assumindo ontologia como caracterização de um domínio e explicitando diferenças e sobreposições com os outros termos.

Uma *taxonomia* é uma hierarquia de termos, na qual podem existir diferentes tipos de relação pai-filho (parte-todo; tipo-instância). Já um *tesauro* pode ser considerado uma extensão de taxonomia, comportando a inclusão de regras de uso de vocabulário, definições, sinônimos e antônimos. Compreende, portanto, além de relações hierárquicas, relações associativas. Por fim, *ontologias* (as específicas de domínio, pelo menos) são mais detalhadas; podem – e devem – conter mais níveis hierárquicos. Em um tesauro, as relações "termo genérico" / "termo específico" podem significar tanto uma relação de hiperonímia quanto de meronímia. A palavra *cachorro*, por exemplo, está relacionada a *mamífero*, em uma relação de hiperonímia; mas a palavra *cabelo* está relacionada a *corpo*, em uma relação de meronímia. Já os "termos associados" cobrem diversas relações semânticas, sem especificação. Alguns dos termos associados a cachorro são: *au-*

au, bassê, *labrador*, *latido*, *cadela*, *canil*, *cão* e *carrocinha*[7]. Nas ontologias, esta ambigüidade de relações não é possível: isto é, considerando, por exemplo, relações hierárquicas de meronímia e hiponímia, é preciso que haja uma distinção formal entre os dois tipos de relação – e não apenas um rótulo geral "termo geral – termo específico" que as abarque.

Neste trabalho entendo como ontologia um conjunto de termos, associados com definições em linguagem natural, que utiliza relações formais e é relativo a algum domínio de interesse (Hovy, 2002). Em termos gerais, trata-se de uma forma de organização do conhecimento de um domínio – o que também está de acordo com a definição de Gruber. Uma ontologia deve ser capaz de capturar uma série de relações semânticas entre termos, não apenas uma relação de inclusão de classe, como é o caso da relação de hiponímia. Ainda assim, assumindo que, em termos práticos, o resultado final deste trabalho é a elaboração automática de uma taxonomia (pois as relações extraídas são de hiponímia), uma vez que a metodologia adotada não impede a possibilidade de inserção de outros tipos de relação semântica, mantenho o uso de *ontologia*. Mas, reforçando o que foi dito, há consciência de que estou tratando, especificamente, de uma porção da ontologia – estou tratando de uma taxonomia.

2.5. Sobre taxonomias e hipônimos

Tradicionalmente, de um ponto de vista representacionista, é tarefa da semântica lexical representar o significado de cada palavra e explicar as relações sistemáticas entre esses significados (Saeed, 1997). Essas relações entre significados – nas quais a hiponímia se inclui – tomam por base, em termos gerais, a distinção clássica entre propriedades essenciais e acidentais. Especificamente, a estabilidade das relações é garantida pelas propriedades essenciais, imutáveis.

No caso da hiponímia, a inclusão de uma palavra numa classe feita com base em uma propriedade acidental não é considerada hiponímia. Isto é, uma relação como

[7] Exemplos retirados de Thesaurus da Língua Portuguesa do Brasil, disponível em http://alcor.concordia.ca/%7Evjorge/Thesaurus/. Acessado em 19/12/2006

gato < animal de estimação

não é uma relação de hiponímia válida porque toma por base uma propriedade como "*domesticável*", que não faria parte das propriedades essenciais de *gato*. Obviamente, esbarramos aqui na discussão sobre quais seriam as propriedades essenciais.

Esta seção descreve como se comporta a relação de hiponímia e sua conexão com a taxonomia de um ponto de vista tradicional, tomando como principais referências os trabalhos de David Cruse (1986, 2004) e John Lyons (1980).

Uma taxonomia pode ser considerada um tipo de configuração lexical, isto é, um vocabulário pode se estruturar em termos hierárquicos. Esse vocabulário tanto pode ser um vocabulário controlado, específico, que caracteriza as taxonomias científicas, quanto o vocabulário de uma língua natural, que por sua vez caracteriza as chamadas taxonomias populares (Lyons, 1980) ou naturais (Cruse, 1986).

Cruse (1986) e Lyons (1980) apontam uma série de características que distinguem as taxonomias formais (ou científicas) das naturais. Nas taxonomias formais, por exemplo, co-hipônimos devem ser incompatíveis: *melancia* e *abacaxi* são co-hipônimos e incompatíveis. Já nas línguas naturais, dois hipônimos do mesmo superordenado não são, necessariamente, incompatíveis: *romance* e *capa dura* (*hardcover*) são hipônimos de *livros*, mas não são incompatíveis – um *romance* pode ser de *capa dura*. De fato, este tipo de incompatibilidade (que é problema para Cruse (1986), mas não para Cruse (2004)) se deve justamente à imprecisão dos limites dos conceitos, o que será abordado mais adiante. Outra diferença entre taxonomias formais e naturais diz respeito à quantidade de níveis. As taxonomias naturais caracterizam-se por ter, no máximo, cinco níveis de profundidade, e mesmo esse número já é bastante raro. Já as taxonomias técnicas ou científicas não têm um número limitado de níveis. Por fim, considerando, de um ponto de vista formal, uma taxonomia ideal, todos os ramos possuem nós em cada nível. Quanto a isso, taxonomias naturais estão longe do ideal, pois estão repletas de lacunas lexicais (*lexical gaps*), isto é, termos para os quais não há um item superordenado. De fato, o vocabulário de uma língua natural pode estar

estruturado hierarquicamente a partir de diversos pontos de origem. Não há nenhum termo superordenado em relação a todos os outros, mas, segundo Lyons, "é inegável a existência de um certo grau de organização hierárquica em todos os níveis do vocabulário das línguas já estudadas" (1980:243).

Levando em conta as diferenças entre taxonomias naturais e científicas, a taxonomia apresentada neste trabalho é uma taxonomia híbrida: por um lado, apresenta características das taxonomias científicas – é específica de um domínio, e baseada em um corpus que contém textos técnicos. Por outro, as informações fonte para a sua elaboração vêm de textos - isto é, vêm de linguagem natural – o que a aproxima das taxonomias naturais.

A relação de hiponímia é a relação-chave de uma taxonomia. Trata-se de uma relação entre uma palavra mais específica (subordinada) e uma palavra mais geral (superordenada), como a relação entre *melancia* e *fruta*. Certamente a hiponímia pode ser considerada uma das formas mais importantes de estruturação do vocabulário, já que a inclusão dos termos em classes possibilita generalização, que se traduz em economia e aproveitamento de informação.

Do ponto de vista lógico, a hiponímia é caracterizada a partir de três critérios: (i) inclusão de classes; (ii) implicação unilateral; (iii) transitividade.

De acordo com o primeiro critério, tem-se que o item subordinado está incluído na classe do superordenado (ou, de modo inverso, que o item superordenado contém o item subordinado): *melancia* está incluída na classe das *frutas* (ou a classe das *frutas* inclui *melancia*).

O critério (ii), implicação unilateral, explora o fato de que a frase *Maria ganhou uma rosa* implica *Maria ganhou uma flor*, mas *Maria ganhou uma flor* não implica *Maria ganhou uma rosa*.

Por fim, o critério da transitividade, o que mais importa para a ontologia, pois permite a realização de inferências, mostra que hiponímia é uma relação transitiva: se X é hipônimo de Y e Y é hipônimo de Z, então X é hipônimo de Z: *melancia* é uma *fruta*; *fruta* é um *alimento*; então *melancia* é um *alimento*. Porém, como aponta Cruse (2004), há diversos casos em que a cadeia de transitividade parece se quebrar, principalmente se um dos elementos da cadeia não é um elemento prototípico: se um *banco de carro* é um *banco* e um *banco* é um *móvel*, então um *banco de carro* é um *móvel* não parece uma relação aceitável.

Além desses três critérios tradicionalmente utilizados, Cruse (1986) acrescenta os seguintes testes para a identificação da hiponímia:

a) equivalência a uma paráfrase em que o superordenado é modificado por um adjetivo:

Rainha é um monarca feminino

b) ocorrência em determinadas construções como:

Gatos *e outros* animais;

Não há flor *mais* bela *que* a rosa;

Ela gosta de *todas as* frutas, *exceto* manga;

Ela lê livros o dia todo – *principalmente* romances

Porém, logo em seguida, Cruse apresenta contra-exemplos que questionam os critérios propostos. Para o critério (a), como criar uma paráfrase equivalente a, por exemplo, *aranha*? A *aranha é um animal*... Já o critério (b) não oferece garantias de discriminação de hipônimos, pois consideraria como tais os elementos presentes em

(1) Gatos *e outros* animais de estimação;
(2) Cobras *e outras* criaturas venenosas;
(3) *Não há* arma *tão* versátil *quanto* uma faca.

Para Cruse (1986), nenhuma das expressões acima contém itens relacionados por meio de hiponímia, visto que as definições abaixo – que seriam fundamentais para a caracterização da hiponímia – não correspondem à realidade:

(1) ? um gato é necessariamente um animal de estimação
(2) ?uma cobra é necessariamente uma criatura venenosa
(3) ?uma faca é necessariamente uma arma

Como discorda de que os exemplos (1), (2) e (3) sejam exemplos de hiponímia, Cruse (1986) propõe uma subdivisão entre os hipônimos: os taxônimos (*taxonymys*). Taxônimos de um item lexical são um subconjunto de seus hipônimos, e seriam elementos cruciais para uma hierarquia lexical taxonômica – com isso, (1), (2) e (3) não seriam taxônimos, seriam apenas hipônimos. A diferenciação entre hiponímia e taxonímia poderia ser feita por meio do seguinte "contexto diagnóstico":

Um X é um tipo de Y

Se X é um taxônimo de Y, então o resultado é considerado normal:

(4) Um labrador é um tipo de cachorro.
(5) Uma rosa é um tipo de flor
(6) Uma banana é um tipo de fruta.

Nos exemplos acima, X é, de fato, um hipônimo de Y. Porém, nem todos os hipônimos levam a um resultado normal neste contexto:

(7) Uma rainha é um tipo de monarca
(8) Um garçom é um tipo de homem.

Segundo Cruse (1986), esses problemas decorreriam da multiplicidade de contextos em que a expressão "tipo de" é utilizada na linguagem cotidiana. Um dos contextos irrelevantes para a identificação ocorreria, por exemplo, em perguntas com o formato ambíguo como "*Que tipo de pessoa é ela?*" e "*Aquele tipo de pessoa que nunca paga suas contas*". Uma pergunta como "*Que tipo de árvore é aquela?*" provavelmente deseja uma resposta taxonômica. Por outro lado, alguém que pergunta "*Que tipo de árvore você está pensando em colocar no quintal?*" poderia muito bem se satisfazer com uma resposta "*Uma que dê bastante sombra*".

Ainda segundo Cruse (1986), "reconhecer uma taxonímia é uma coisa; *descrever sua essência natural* é uma outra tarefa, mais difícil" (1986:139) (grifo meu)[8]. Porém, ao invés de abandonar a distinção entre taxônimos e hipônimos, o autor segue tentando dissecar as diferenças entre as duas categorias, apoiando-se em algumas abordagens que, segundo ele, parecem iluminar a questão[9].

A primeira dessas abordagens diz respeito a uma forte correlação entre taxônimos e os chamados "tipos naturais" (*natural kind terms*), por um lado, e entre hipônimos não-taxonômicos e "tipos nominais" (*nominal kind terms*), por

[8] "Recognizing a taxonomy is one thing; describing its essential nature is another and more difficult task" (Cruse, 1986:139)

[9] "However, there are two or three lines of approach which seem to throw some light on the matter" (Cruse, 1986:139-140).

outro[10]. Porém, o próprio Cruse apresenta como contra-exemplo para essa distinção a taxonomia para "cores de cabelo" (hair-colour), uma taxonomia bem formada e que é baseada em tipos nominais[11].

A segunda consideração referente à caracterização da relação de taxonímia seria em termos de prototipicidade: categorias taxonômicas seriam constituídas por elementos prototípicos. Neste caso, poderíamos afirmar que a divisão taxonômica entre *garanhões* e *éguas* é taxonomicamente anômala, pois o critério *sexo* não seria o melhor critério para diferenciação de categorias. Mas Cruse também refuta esse argumento, afirmando que nem sempre a divisão em classes com base em prototipicidade é possível.

Cruse (1986) chega então à conclusão ("pessimista para a teoria semântica", segundo ele) de que

> "Perhaps (...) there are no invariable principles to be applied which inevitably lead to unique taxonomies; perhaps we merely seek to create the closest analogues we can to natural species. Exactly how close we get will of course depend on the nature of the category being sub-divided"
>
> (Cruse, 1986:144)

Finalmente, em Cruse (2004) a tentativa de distinção entre hipônimos e taxônimos é abandonada. Em nome do que chama "dynamic construal approach", o autor propõe um questionamento da assumida estabilidade do significado das palavras:

> "There are many different approaches to the study of semantic properties of words, but most of them take it for granted that each Word has a stable, inherent attribute called a 'meaning', which it is the job of a lexical semanticist to describe. (…) there is a general agreement that word meanings exist, and that logical and structural aspects of meaning, such as sense relations, and certain logical properties of utterances, are either directly represented in the lexical entry or can be inferred from the lexical entry.
> Of course, there is also general agreement that meaning is highly context-dependent (…).
> However, it has proved extremely difficult to achieve a satisfactory picture using these assumptions, and some linguists have begun to explore the consequences of abandoning the assumption of stable word meanings. (…)Well, the proposal is not

[10] Tipos naturais referem-se a classes de entidades que existem na natureza, como árvores, gatos, banana; e tipos nominais referem-se a agrupamentos mais arbitrários.

[11] Exemplo da taxonomia:
hair-colour>blonde, red-head, brunette;
blonde>ash blonde, strawberry-blonde

> that words have no stable semantic properties, but rather that these properties are not meanings"
>
> (Cruse, 2004: 261-262)

Nesta nova abordagem, palavras não têm significados que lhes são permanentemente atribuídos. Os significados emergem do uso como o resultado de vários processos interpretativos. O que as palavras possuem como uma propriedade permanente é um mapeamento entre um corpo de conteúdos conceituais (que Cruse chama de "*purport*"), que é parte essencial da matéria prima necessária ao processo de construção da interpretação (*construal process*), mas que sub-determina quaisquer significados específicos. Esses processos de interpretação que resultam em significados contextualizados estão sujeitos a restrições de vários tipos e com diferentes forças, o que torna determinadas leituras mais plausíveis que outras.

Esta nova abordagem de Cruse parece mais compatível com o enquadre wittgensteiniano assumido aqui:

> "it is an essential feature of the dynamic construal approach that, just as words are not associated with specific meanings, nor are they associated with specific conceptual categories, but with bodies of purport which allow variable construal in different contexts."
>
> (Cruse, 2004:267)

Assim, do mesmo modo que não podemos concordar com a abordagem de Cruse (1986) quando desconsidera relações de hiponímia como *gatos são animais de estimação*, concordamos com Cruse (2004) quando afirma que

> "Taking the dynamic construal view (...): *cat* is a hyponym of *pet* in *cats and other pets*, but not in *It's a cat, therefore it is a pet*, and the difference is due to the fact that the construed categories are different in the two contexts."
>
> (Cruse, 2004:268)

Porém, se tanto para a abordagem pragmática wittgensteiniana quanto para a abordagem de Cruse (2004) não há dificuldade quanto à aceitação de determinados termos em determinadas categorias, desde que possíveis em um contexto de uso, e ainda que a inclusão na classe seja pouco convencional, tanta "permissividade", mesmo que teoricamente motivada, leva a um problema significativo no âmbito do PLN: como então proceder a uma avaliação dos

resultados, isto é, como será possível saber se a identificação automática de relações de hiponímia é realmente eficaz, se o processo de inferências produziu conhecimento correto e como comparar esses resultados com o de outros trabalhos semelhantes? O capítulo 3 problematiza a questão da avaliação neste tipo de pesquisa.

3
Critérios para a elaboração e avaliação de ontologias

A dificuldade em se chegar a um consenso sobre o que seja uma ontologia tem reflexos tanto no que concerne aos critérios metodológicos que devem ser observados no momento da sua elaboração, quanto aos critérios de avaliação e comparação das ontologias construídas.

Ontologias são objeto de diferentes áreas e, com isso, apresentam, freqüentemente, diferentes objetivos. E, como afirma Gruber, o importante é para que a ontologia serve – o que importa é sua função, e, portanto, diferentes diretrizes irão guiar sua elaboração.

Nesta seção, descrevo duas propostas gerais para a elaboração de ontologias, que correspondem a duas visões distintas sobre o problema: a proposta de metodologia de Hovy (2005), compatível com o que chamei no capítulo 2 de visão tradicional de ontologias e significados, e a proposta de metodologia de Brewster e Wilks (2004), compatível com o que chamei de perspectiva relativista. Em ambas as propostas, trata-se de metodologias teoricamente motivadas que devem direcionar o "construtor de ontologias" e facilitar a consistência e a exatidão, em todas as etapas da elaboração.

3.1.1.
Critérios para a elaboração de ontologias "tradicionais"

Com base na experiência própria de coordenador de projetos relacionados ontologias, Hovy (2005) distingue cinco diferentes motivações que irão se refletir em diferentes abordagens para a elaboração de ontologias: abordagem da filosofia, da ciência cognitiva, da lingüística, da inteligência artificial e da computação. Assumindo ser a principal tarefa do "ontologista" a discussão sobre a criação de um novo termo e sua localização na ontologia com relação aos outros termos já existentes, para uma posterior especificação adicional e definição, Hovy descreve como todo este processo decisório poderia ocorrer, levando em consideração a

"personalidade" do ontologista – se filósofo, se cientista cognitivo, se lingüista, etc.

Especificamente, a proposta de metodologia de Hovy (2005), que pode ser chamada de *refinamento gradual contínuo*, é composta por 7 etapas:

1) Determinação das características gerais da ontologia que se quer construir: domínio, objetivo, nível desejado de granularidade, antecedentes teóricos e conceituais, público-alvo etc;

2) Coleta de todas as fontes de conhecimento adicionais, incluindo ontologias anteriores, estruturas de alto nível, glossários de termos, algoritmos e ferramentas, descrições teóricas já existentes, etc;

3) Delimitação do principal fenômeno sob consideração: identificação dos conceitos nucleares, tipos ou características permitidos, etc. Começar com uma ontologia de alto nível, já existente, pode ser útil;

4) Listagem de todos os termos/conceitos importantes para a tarefa. Os termos podem ser derivados de algum modelo de (meta)dados, de algum algoritmo do sistema, de relatórios de especialistas, etc;

5) Registro explícito, para cada conceito, dos princípios e fatores que justificam a sua criação (ainda que a definição seja incompleta ou informal, desde que contenha as principais características de interesse). Identificação e definição das relações e dos conceitos;

6) Inspeção da ontologia inicial, buscando corrigir irregularidades, desequilíbrios, etc;

7) caracterização da ontologia quando pronta, registrando seus parâmetros essenciais (cf. descrição detalhada em Hovy, 2002).

Como o próprio Hovy assume, a utilização dessa metodologia leva tempo e exige esforço. Nem todos os aspectos se aplicam a todos os casos, e nem a todos os domínios ou "estilos de ontologização".

3.1.2.
Critérios para a elaboração de ontologias baseadas em corpus

Outra sugestão de metodologia para a elaboração de ontologias é apresentada em Brewster e Wilks (2004). A principal diferença desta abordagem é o reconhecimento de que "*given the `info-smog' we live in, hand-crafting is impractical and undesirable*" (Brewster e Wilks, 2004:4). Os autores admitem, no entanto, que embora a construção totalmente automática de ontologias ainda seja um desafio, as atuais ferramentas de PLN possibilitam em grande parte a automação da tarefa, reduzindo significativamente o trabalho manual.

Brewster e Wilks (2004) apresentam uma série de critérios metodológicos com o objetivo de ajudar na escolha das ferramentas adequadas para a construção de taxonomias/ontologias[12]. São eles:

1) *Coerência*: A taxonomia deve ser, para o usuário, uma organização coerente, de bom-senso, dos conceitos ou termos. A coerência depende de os termos e as associações entre eles integrarem a "conceitualização compartilhada" a que Gruber (1993) se refere. A noção de coerência é dependente da aplicação.

Brewster e Wilks (2004) ressaltam que, com essa escolha, torna-se muito difícil avaliar qualitativamente uma taxonomia ou hierarquia, pois não haveria um critério totalmente estabelecido capaz de decidir se uma taxonomia é correta, ou se uma é melhor que outra. Diferentemente de áreas como a recuperação de informação, em que as medidas de precisão e recuperação são amplamente aceitas, não há medidas equivalentes para ontologias, já que o conhecimento não é uma entidade quantitativa.

2) *Herança múltipla*: Esta característica, já apontada por Noy e McGuiness (2001), diz respeito à localização de um termo em múltiplas posições na taxonomia. Freqüentemente, um único termo apresenta uma variedade de facetas que justifica sua localização múltipla em uma taxonomia ou ontologia. O termo

[12] Embora os autores apresentem os critérios como norteadores para a elaboração de ontologias, tais critérios também podem ser utilizados em um momento posterior, de avaliação das ontologias.

"tuberculose pulmonar", por exemplo, é um distúrbio respiratório e, igualmente, um "distúrbio infeccioso".

3) *Facilidade na computação*: Como um dos problemas das bases de conhecimento em geral é a manutenção/ atualização, é importante que o método escolhido não apresente uma grande complexidade computacional, de modo a não tornar sua atualização muito custosa.

4) *Rótulos únicos*: É importante que todos os nós tenham rótulos únicos, ainda que sejam compostos por mais de uma palavra. Grupos de palavras caracterizados por uma única etiqueta são mais facilmente compreendidos pelo usuário. Os autores citam o exemplo dos rótulos múltiplos de Sanderson e Croft (1999), em que um grupo de documentos é caracterizado pelo seguinte conjunto de termos: *bateria Califórnia tecnologia milha estado recarga impacto oficial custo hora governo*, o que dificulta sua categorização / identificação.

5) *Fonte de dados*: Brewster e Wilks sugerem que os dados para a construção da ontologia devam vir de duas fontes: a) documentos (fontes primárias); b) dados provenientes de uma taxonomia já existente ("seed taxonomy"), que funcionariam como uma estrutura de dados para revisar ou para servir de esboço inicial, quando necessário.

3.2. Formas de avaliação de ontologias

Boa parte dos trabalhos em PLN utiliza como forma de avaliação as medidas de precisão e abrangência. Tais medidas são tradicionalmente empregadas na área de Recuperação de Informação, e baseiam-se na noção de relevância. A precisão é a proporção de documentos recuperados em uma busca que são relevantes para o usuário. A abrangência corresponde à proporção de documentos relevantes que foram recuperados[13]. Para a realização desses cálculos,

[13] Precisão = $\frac{\text{número de documentos relevantes identificados}}{\text{número total de documentos identificados}}$

Abrangência = $\frac{\text{número de documentos relevantes identificados}}{\text{número de documentos relevantes na coleção}}$

é fundamental a existência de material já corretamente identificado, isto é, é fundamental a existência de um "gabarito", com os quais os novos resultados da identificação automática possam ser comparados – esse "gabarito" pode ser um corpus etiquetado, no caso de tarefas de PLN, ou , no caso da extração de informações, uma coleção de consultas, documentos e julgamentos de relevância conhecidos.

Na avaliação de ontologias (e taxonomias e léxicos semânticos), porém, as noções de precisão e abrangência não são facilmente aplicáveis, pois a natureza da tarefa é diferente. A medida de precisão poderia refletir a quantidade de conhecimento corretamente identificado na ontologia, com relação a todo o conhecimento disponível na ontologia. A abrangência poderia refletir a quantidade de conhecimento corretamente identificado com relação a todo o conhecimento que deveria ser identificado. O problema está em como definir "o conhecimento que deve ser adquirido", já que o mesmo conjunto de fatos pode levar a diferentes interpretações e, conseqüentemente, a diferentes tipos de "conhecimento" (Brewster et al., 2004).

De fato, a avaliação de ontologias e taxonomias vem sendo bastante discutida e ainda não existem abordagens abrangentes e gerais para o problema. Como afirmam Brewster et al.(2004), "*There are inherent problems in trying to evaluate an ontology as it is not clear what exactly one is trying to evaluate*" (2004:1). De uma maneira geral, para os autores, boas ontologias são aquelas que servem aos seus propósitos.

Alguns trabalhos sugerem que a avaliação de ontologias deve ser feita numa comparação com um modelo ideal (*golden model*), como é o caso de Hovy (2002) e Maedche e Staab (2002). Apesar de trazer a vantagem de possibilitar a utilização de medidas conhecidas como precisão e abrangência, este tipo de comparação, porém, não será abordado neste trabalho porque não temos um modelo ideal de ontologias para o português (e sua elaboração é o processo custoso que justamente se quer evitar), e porque a própria idéia de "modelo ideal" parece não se ajustar ao modelo apresentado aqui. Além disso, quando os resultados diferem dos do modelo ideal, é difícil detectar a origem do problema: se o corpus é inapropriado, se a metodologia é inadequada ou se há uma diferença entre o conhecimento presente no corpus e o modelo ideal (Brewster et al., 2004). Ainda na linha "modelo ideal", uma possibilidade de avaliação é comparar os

resultados obtidos com as relações presentes na WordNet (Fellbaum, 1998), como fazem alguns trabalhos (Snow et al., 2005; Widdows, 2003; Lin e Pantel, 2002). Porém, considerando os limites da Wordnet com relação a termos específicos de domínio, esta proposta também parece inadequada.

Um outro tipo de avaliação é apresentado por Velardi et al. (2005), que sugerem a comparação entre os resultados da ontologia automática e a avaliação humana. Para os autores, um dos objetivos da avaliação das ontologias geradas automaticamente não é apenas a comparação das diferentes abordagens, mas a verificação da capacidade de um dado processo automático de competir com o processo tipicamente humano de conceitualização de um determinado domínio. Para Verlardi et al., portanto, as questões fundamentais que se apresentam são (i) poderia um método automático simular esse processo humano? e (ii) é possível oferecer a especialistas formas de mensurar a adequação de um conjunto de conceitos como modelo de um domínio?

Além disso, como freqüentemente especialistas têm dificuldades para avaliar o conteúdo formal de uma ontologia computacional, sua função é comparar a intuição que têm do domínio com a descrição deste fornecida pela ontologia.

Tentando minimizar estas dificuldades, Velardi et al. desenvolveram um método para a geração automática de glosas a partir das relações entre os conceitos da ontologia. Com isso, eliminam a dificuldade dos especialistas de lidar com os aspectos formais do conteúdo – glosas oferecem uma descrição, em linguagem natural, das especificações formais atribuídas aos conceitos da ontologia[14]. A comparação entre intuição e glosas é, sem dúvida, um trabalho mais simples.

De qualquer maneira, contudo, a tarefa de avaliação é manual, um trabalho custoso e subjetivo, dependente da intuição do especialista – o rigor da avaliação é de difícil verificação. Porém, a idéia de criação automática de glosas a partir das definições parece bastante interessante não apenas para a avaliação por humanos, mas em outros contextos que envolvam geração automática de textos.

[14] Exemplo de glosa gerada, no domínio turismo:
termo: *affiliated_hotel*
glosa: "*a kind of hotel, a building where travelers can pay for lodging and meals and other services, being joined in close association*".

O modelo de avaliação de Brewster et al. (2004) é o que mais combina com a proposta apresentada neste trabalho. Compartilhando os pressupostos de que ontologias devem ser baseadas em corpus, Brewster et al. propõem uma avaliação direcionada aos dados (*data-driven*). Sugerem, especificamente, uma medida para a adequação entre a ontologia e o domínio de conhecimento (*ontological fit*).

Deste modo, se corpora devem ser a fonte mais efetiva de informação para a construção de ontologias (ou de uma grande porção da ontologia, como é sugerido em Brewster et al., 2001), a avaliação deve consistir na identificação da adequação da ontologia com o corpus. Para tanto, uma sugestão simples de Brewster et al. (2004) é a extração automática de termos do corpus e a contagem do número de termos que estão, simultaneamente, no corpus e na ontologia. A ontologia seria "penalizada" para termos presentes no corpus e ausentes na ontologia, e para termos presentes na ontologia e ausentes no corpus. A proposta dos autores, no entanto, é um pouco mais sofisticada: uma metodologia em 3 etapas, que dificilmente poderíamos aplicar por ser dependente de WordNet e de um corpus semanticamente anotado, ferramentas das quais não dispomos em português.

Por fim, descrevo brevemente a proposta de Etzioni et al. (2005) que, embora não seja aplicada a ontologias, diz respeito à avaliação de extração de relações de hiperonímia em um corpus não-anotado – prescindindo, portanto, das tradicionais medidas de precisão e abrangência.

Etzioni et al. apresentam um sistema – KnowItAll – de aquisição automática de instâncias (nomes próprios) e de classes dos domínios Turismo/Geografia, Filmes e Cientistas, que toma como fonte a Web.

Para avaliar a extração das instâncias, um módulo do sistema, chamado "avaliador", calcula a PMI (*Pointwise Mutual Information*) entre a instância extraída e "sintagmas discriminadores" associados à classe em questão. Esses sintagmas discriminadores podem ser "X é um Y", ou "Y X" em que X é a instância e Y a classe-alvo. Por exemplo, para as instâncias "Brasil" e "O Invasor" (X) e as classes-alvo "país" e "filme" (Y), seriam sintagmas discriminadores construções como

"*Brasil é um país*" ou "*o país Brasil*";

"*O Invasor é um filme*" ou "*o filme O Invasor*"

O escore PMI calculado é o número de respostas (hits) para uma busca que combina o sintagma discriminador e a instância, dividido pelo número de hits para a instância sozinha:

$$PMI = \frac{| \text{Hits (Brasil é um país)} |}{| \text{Hits (Brasil)} |}$$

Porém, como destacam os autores, o score PMI bruto é, tipicamente, uma fração muito pequena, mesmo para instâncias positivas. Além disso, o cálculo não dá a probabilidade de uma instância ser membro da classe, apenas a probabilidade de ver o sintagma discriminador em *webpages* que contêm a instância.

Embora satisfeitos com suas medidas de avaliação, Etzoini et al. alertam para os seguintes problemas:

- dados esparsos (mesmo usando a web como corpus);
- polissemia: *Botafogo* pode ser um bairro e um clube de futebol. Quando o sentido em que se está interessado é o menos freqüente, os scores tendem a ser baixos.

Como, neste trabalho, utilizo um corpus de dimensões mínimas, se comparado à web (cf. capítulo 5), fica claro que o problema da escassez de dados irá se repetir. Por outro lado, a utilização de "sintagmas discriminadores" parece uma boa idéia para testar a validade das relações extraídas automaticamente do corpus. Assim, para a frase (1) abaixo

(1) *..., transtorno obsessivo-compulsivo, distimia , transtorno afetivo bipolar, abuso de álcool e outras substâncias*

em que são extraídas as relações

transtorno obsessivo-compulsivo < substância;
distimia < substância;
transtorno afetivo bipolar < substância;
abuso de álcool< substância,

uma forma de avaliação seria a realização de uma pesquisa na internet utilizando como expressão de busca as expressões "*abuso de álcool é uma substância*"; "*distimia é uma substância*", ou "*substância abuso de álcool*" ou "*substância*

distimia". Como em todas as expressões o resultado da busca é zero, as relações são descartadas.

Porém, nem sempre os resultados dos padrões discriminadores serão confiáveis. Na frase (2)

> *(2) O grupo étnico materno foi definido por branco, pardo e negro, considerando critérios como cor da pele, textura do cabelo e formato do nariz.*

são extraídas as relações

cor da pele < critérios
textura do cabelo < critérios
formato do nariz < critérios

Em seguida, uma pesquisa na internet usando os "padrões discriminadores" obtém os seguintes resultados (tabela 1):

Padrão discriminador	Qtde de documentos recuperados
"cor da pele é um critério"	1
"critério cor"	89
"critério cor da pele"	2
"textura do cabelo é um critério"	Zero
"textura é um critério"	Zero
"critério textura"	Zero
"critério formato"	3
"critério formato do nariz"	Zero
"formato é um critério"	Zero
"formato do nariz é um critério"	Zero

Tabela 1: Resultados de busca na Internet por padrão discriminador

Outro problema na utilização dos "padrões discriminadores" está na estrutura morfossintática do sintagma utilizado na busca. O fato de mecanismos de busca não considerarem acentos é mais um complicador para a avaliação em língua portuguesa, pois, a uma busca como "*textura é um critério*", também correspondem resultados como "*textura e um critério*". Outro problema é que é preciso fazer a concordância de gênero ("*é um*" / "*é uma*"), complicador que não aparece para a língua inglesa ("*is a*").

Neste capítulo, foram apresentados critérios para a elaboração e avaliação de ontologias. Quanto à elaboração, existem basicamente duas maneiras de enfrentar o problema:

- assumir que ontologias devem ser elaboradas sobretudo de forma manual, em um trabalho que visa, principalmente, a inserção de conceitos relevantes do domínio em questão em uma estrutura formal;
- utilizar critérios mais gerais de elaboração, que levem em conta simultaneamente características desejáveis do ponto de vista da caracterização do domínio e da automação do processo.

Em ambas as propostas está prevista a intervenção humana, seja de maneira explícita, em todas as etapas do processo, como sugere Hovy (2005), seja de maneira implícita e minimizada, como na incorporação de uma taxonomia pré-existente ("seed taxonomy"), que pode funcionar como um esboço de ontologia, como sugerem Brewster e Wilks (2004).

Uma conseqüência da dificuldade em se atingir um consenso sobre metodologia para a elaboração de ontologias é a ausência de critérios sistemáticos para sua avaliação.

Em geral, as avaliações são sustentadas por dois diferentes paradigmas:

- A comparação com algum modelo ideal, que pode ser uma outra ontologia ou a WordNet;
- A avaliação manual por especialistas.

Porém, ao se considerar especificamente a elaboração de ontologias a partir de corpus, evidencia-se a necessidade de uma outra forma de avaliação – uma avaliação que meça a adequação entre a ontologia e o corpus, como sugerem Brewster et al. (2004).

4
Trabalhos relacionados à extração automática de hiperonímia

Neste capítulo, relato os principais trabalhos que tratam da extração automática de relações de hiperonímia a partir de textos. Começo com uma apresentação detalhada das wordnets que, embora não sejam elaboradas automaticamente, são freqüentemente vistas como um modelo a ser atingido.

4.1.
WordNet, EuroWordNet e Wordnet.Br

A WordNet (Fellbaum, 1998) é um léxico semântico relacional desenvolvido para a língua inglesa, disponível para uso online[15], cujos principais objetivos são (i) oferecer uma combinação de dicionário e tesauro que seja mais utilizável de um ponto de vista intuitivo; e (ii) dar suporte a tarefas que envolvem a análise automática de textos. Na WordNet, as palavras estão agrupadas em conjuntos de sinônimos chamados *synsets*. Sua forma de organização se baseia nos resultados de experimentos psicolingüísticos e busca reproduzir a estrutura do nosso léxico mental.

A WordNet distingue entre substantivos, verbos, adjetivos e advérbios. Cada *synset* contém um grupo de palavras ou expressões sinônimas. A maioria dos *synsets* se conecta a outros *synsets* por meio de relações semânticas como hiperonímia, hiponímia, coordenação, holonímia e meronímia (substantivos); hiperonímia, troponímia e acarretamento (verbos); substantivos relacionados e formas de particípio de verbos (adjetivos); adjetivos-raiz (advérbios).

As relações de hiperonímia/hiponímia entre os *synsets* de substantivos podem ser compreendidas como relações entre categorias conceituais, o que permite que a WordNet seja interpretada (e utilizada) como uma ontologia lexical

[15] A WordNet está disponível em http://wordnet.princeton.edu/.

na IA. No nível mais alto, as hierarquias estão organizadas em 25 primitivos nominais.

Atualmente (2006), a WordNet, "a mãe de todas as wordnets" (Fellbaum, 1998), conta com mais de 150.000 palavras organizadas em mais de 115.000 synsets, constituindo-se em um modelo para outras línguas e tornando-se um dos recursos de maior impacto no PLN. A WordNet vem se desenvolvendo desde 1985 na universidade de Princeton, em um projeto que já recebeu mais de 3 milhões de dólares para seu desenvolvimento.

Tomando como modelo a WordNet de Princeton, foi desenvolvida a EuroWordNet (Vossen, 1998), uma base de dados multilingüe que integra wordnets de diversas línguas européias. A diferença fundamental entre a WordNet de Princeton e a EuroWordNet é o fato de a segunda ser multilingüe – as wordnets das diversas línguas são relacionadas pelo "Inter-Lingual-Index" (ILI), uma lista de *synsets* que corresponde aos *synsets* da WordNet de Princeton.

A Wordnet.Br é a versão brasileira da WordNet, que conta atualmente com cerca de 11.000 verbos, 17.000 substantivos, 15.000 adjetivos e 1.000 advérbios, num total de 44.000 palavras e 18.500 *synsets* (Dias-da-Silva et al., 2006). Para sua elaboração, a Wordnet.Br reaproveita material disponível em outras fontes, como as versões eletrônicas dos dicionários *Aurélio* e *Michaelis*, dicionários de sinônimos e antônimos, um dicionário analógico e um dicionário de verbos do português. Porém, a maior parte do trabalho de elaboração é feita manualmente (Dias-da-Silva et al., 2006). Na fase atual de desenvolvimento, os lingüistas que participam do projeto têm realizado (i) a análise da consistência semântica dos *synsets*; (ii) a coleta e seleção das frases-exemplo, extraídas de corpus.

4.2. Extração automática de hiperonímia

Condamines e Rebeyrolle (2000) classificam em métodos top-down ou bottom-up as diversas técnicas desenvolvidas para a extração automática de relações semânticas a partir de textos. Métodos top-down utilizam padrões lingüísticos pré-definidos; as técnicas para a aquisição de relações semânticas se baseiam em regras criadas manualmente para a extração dos dados. O trabalho de Hearst (1992, 1998) se enquadra nesta abordagem – e a desvantagem da técnica

consiste justamente na tarefa manual de codificação das regras, que pode requerer um grande trabalho. Nos métodos bottom-up não é fornecida nenhuma informação sobre os dados que serão extraídos. As palavras são agrupadas (ou classificadas) por meio de técnicas de agrupamento (clusterização) que se baseiam na similaridade entre contextos de palavras. De maneira geral, o problema desta abordagem é que freqüentemente os grupos de palavras (clusters) não são rotulados – trata-se de aglomerados semânticos, o que pode ser um problema para determinadas aplicações. Freqüentemente, essa técnica é utilizada na extração de associações entre palavras (Lin e Pantel, 2002; Widdows 2003) e, de maneira mais rara, na elaboração de tesauros (Kilgarriff, 2003). Alguns trabalhos apresentam uma combinação de abordagens top-down e bottom-up, conjugando técnicas de clusterização e codificação de regras (Caraballo, 1999; Cerderberg e Widdows, 2003; Snow et al., 2005; Morin e Jacquemin, 2004).

4.2.1. Os padrões de Marti Hearst

Marti Hearst (1992, 1998) foi a primeira a utilizar a idéia de que determinados padrões léxico-sintáticos poderiam, sistematicamente, expressar determinadas relações semânticas.

Nesse contexto, relações de hiponímia seriam especialmente úteis às tarefas de PLN porque permitiriam a expansão de léxicos existentes, como a WordNet. Com isso, um dos objetivos da metodologia é auxiliar, de maneira automática ou semi-automática, o trabalho de lexicógrafos e construtores de bases de conhecimento dependentes de domínio.

Especificamente, Hearst (1992, 1998)[16] propõe métodos de extração automática de relações léxico-sintáticas e compara os resultados obtidos automaticamente com os obtidos manualmente pela equipe de lexicógrafos da WordNet.

Hearst propõe a identificação, no corpus, de padrões léxico-sintáticos que codifiquem a relação de hiperonímia na língua inglesa e que obedeçam aos seguintes critérios:

- Ocorrência freqüente e em diferentes tipos de texto;
- Indicação (quase) sempre constante da relação de interesse;
- Pouca ou nenhuma necessidade de conhecimento pré-codificado.

Seguindo esses critérios, os padrões encontrados para o inglês foram:

(i) NP_0 such as NP_1 {, NP_2 ... , (and | or) NP_i}
(ii) such NP_0 as {NP ,}* {(and | or)} NP
(iii) NP {, NP}* {,} or other NP_0
(iv) NP {, NP}* {,} and other NP_0
(v) NP_0 {,} including { NP ,}* {or | and} NP
(vi) NP_0 {,} especially { NP ,}* {or | and} NP

onde NP_0 corresponde a um sintagma nominal (SN) hiperônimo e os demais NPs (NP_1, NP_2...NP_i) a SNs hipônimos:

$$SN_0 > SN_1, SN_2, SN_3 ... SN_i$$

Os padrões (i), (iii) e (iii) foram descobertos manualmente, por meio de observação no corpus. Porém, para que a abordagem seja mais abrangente, Hearst sugere um procedimento-padrão de descoberta, por meio do qual os demais padrões foram identificados, e que consiste basicamente de 4 etapas:

- decidir qual a relação lexical de interesse;
- derivar, por meio da WordNet, uma lista de pares de palavra na qual a relação esteja expressa: por exemplo, para a relação de meronímia, o par *carro-volante;*
- extrair sentenças do corpus em que ambas as palavras (*carro* e *volante*) apareçam, registrando o contexto lexical e sintático em que foram encontradas;
- encontrar semelhanças entre esses contextos e tentar generalizar: contextos comuns levam a padrões que indicam a relação de interesse.

A partir desses padrões, quando uma relação de hiponímia é descoberta, o SN encontrado é considerado uma unidade atômica, indivisível. São retirados apenas o que Hearst chama de "modificadores indesejados", como alguns adjetivos comparativos ("*smaller*", "*important*"). Um problema já observado por

[16] A principal diferença entre os dois trabalhos está no corpus utilizado: em 1992, os padrões foram extraídos de *Grolier's Encyclopaedia*; em 1998, de seis meses do jornal *New York Times*.

Hearst, e que produz erros também para a língua portuguesa (cf. seção 6.1) diz respeito à determinação do referente de um sintagma preposicional (SPrep). Para o inglês, Hearst nota que, na maioria das vezes, o substantivo final no SPrep que precede o "*such as*" (no padrão (i)) é o hiperônimo da relação, como no exemplo (1), embora existam inúmeras exceções, como ilustra a frase (2):

*(1) Agar is **[**a substance prepared from a mixture of red algae**]**, such as Geldium, for laborary or industrial use.*
*(2) A bearing is a structure that supports a **[**rotating part of a machine**]**, such as shaft, axle, spindle, or wheel.*

Isto é, para as frases (1) e (2) seriam extraídas, respectivamente, as relações (1') e (2'), em que a relação (1') está errada pois o sintagma hiperônimo é apenas *red algae*. Já as relações extraídas em (2') estão corretas.

(1´) Geldiu < substance prepared from a mixture of red algae

(2') shaft < rotating part of a machine
axle < rotating part of a machine
spindle < rotating part of a machine
wheel < rotating part of a machine

Com relação à ambigüidade do SPrep nos outros padrões, Hearst comenta apenas que, no padrão "*and other*", diferentemente do "*such as*", freqüentemente o SN completo corresponde ao hiperônimo (3), o que ilustraria a dificuldade de se trabalhar com textos, principalmente de jornais, por sua diversidade, em contraste com as estruturas textuais relativamente previsíveis de dicionários e enciclopédias. Como resposta a essas dificuldades, Hearst sugere que uma solução simples seria descartar as orações em que a ambigüidade é possível, buscando-se apenas SNs simples.

(3) *Temples, treasuries, and other important**[**civic buildings**]**.*

Como um dos objetivos de seu trabalho é, automaticamente, aumentar as relações da WordNet, a análise dos resultados é feita por meio de uma comparação entre as relações identificadas automaticamente e as relações de hiperonímia presentes na WordNet. Em geral, Hearst observa que as relações obtidas a partir do corpus de jornal tendem a ser menos taxonômicas, ou prototípicas, do que as encontradas em textos enciclopédicos; são mais

influenciadas pelo contexto em que aparecem, e refletem de forma mais sistemática julgamentos subjetivos e usos metafóricos do que afirmações estabelecidas que constam de enciclopédias. Como exemplo, uma afirmação como "*Casablanca* é um *clássico*" pode ser considerada decorrente de um julgamento de valor (embora Hearst reconheça que enciclopédias muitas vezes afirmam que determinados atores são estrelas, o que não é tão diferente). Do mesmo modo, a declaração "*AIDS* é um *desastre*" pode ser entendida mais como uma relação metafórica do que taxonômica.

Além disso, como a maioria dos termos da WordNet são nomes sem modificadores ou nomes com um único modificador, os algoritmos de Hearst extraem apenas relações que consistem de nomes sem modificadores, tanto no sintagma hiperônimo quanto no hipônimo. A utilidade dessa restrição estaria na dificuldade de se encontrar um procedimento transparente capaz de determinar quais modificadores são importantes. Acrescente-se a isso que, para fins de avaliação, na maioria dos casos é mais fácil julgar a correção de uma relação com substantivos sem modificadores.

No trabalho, 200 instâncias do padrão "*e outros*" foram avaliadas manualmente. Os avaliadores deveriam classificar os resultados de acordo com oito categorias, como mostra a tabela 2, retirada de Hearst (1998):

Freqüência	**Explicação**
38	Alguma versão dos SNs e sua relação correspondente foi encontrada na WordNet
31	A relação não apareceu na WordNet e foi considerada uma relação ótima (em alguns casos ambos os SNs estavam presentes, em outros casos não)
35	A relação não apareceu na WordNet e foi considerada uma relação pelo menos boa (em alguns casos ambos os SNs estavam presentes, em outros casos não)
19	Relação muito geral
8	Relação muito subjetiva, ou que continha referentes inapropriados (e.g., "these")
34	Os SNs envolvidos eram muito longos, muito específicos e/ou muito dependentes de contexto
12	As relações eram repetições dos casos acima
22	As frases não continham a forma sintática apropriada (e.g., "all of the above, none of the above, or other")

Tabela 2: Resultado da avaliação de 200 frases com o padrão "e outros" (Hearst, 1998)

Consciente do alto grau de subjetividade deste tipo de avaliação, e assumindo uma abordagem "cautelosa" na avaliação, 63% das relações extraídas foram consideradas corretas, isto é, passíveis de serem inseridas na WordNet.

4.2.2. Outros trabalhos

Morin e Jacquemin (2004) apresentam um sistema – *Prométhée* – que extrai e utiliza padrões léxico-sintáticos no estilo Hearst a partir de corpus. O processo de extração automática de padrões é realizado em sete etapas:

(a) seleção manual da relação semântica que se deseja identificar;
(b) coleta de uma lista de pares de termos que participam na relação. Esses pares podem ser extraídos de um tesauro, de uma base de conhecimento ou ainda especificados manualmente;
(c) descoberta de frases em que os pares de termos ocorram – as frases são representadas como expressões léxico-sintáticas;
(d) descoberta de contextos comuns que generalizem as expressões léxico-sintáticas – estes contextos são calculados utilizando funções de similaridades e processos de generalização;
(e) Validação dos padrões por um especialista;
(f) Uso dos padrões validados para a extração de outros pares de termos;
(g) Validação dos pares candidatos por um especialista.

De um conjunto inicial, criado manualmente, de 40 pares de termos relacionados por hiperonímia, o sistema Prométhée identificou 11 padrões léxico-sintáticos[17] que consistem de pequenas variações dos padrões identificados em Hearst (1992, 1998). Os padrões estão descritos abaixo, e SN_1 corresponde ao SN hiperônimo e:

(1) {deux | trois...} SN_1 (Lista de SNs)
(2) {certain | quelque | de autre...} SN_1 (Lista de SNs)
(3) { deux | trois...} SN_1 : (Lista de SNs)
(4) { certain | quelque | de autre...} SN_1 : (Lista de SNs)
(5) SN_1 tel que Lista de SN

---

(1) [17] {dois | três...} SN_1 (Lista de SNs)
(2) {certos | alguns | outros...} SN_1 (Lista de SNs)
(3) { dois | três...} SN_1 : (Lista de SNs)
(4) {certos | alguns | outros...} SN_1 : (Lista de SNs)
(5) SN_1 tais como Lista de SN
(6) SN_1, particularmente SN_2
(7) SN_1 como Lista de SNs
(8) SN_1 tais como Lista de SNs
(9) SN_2 {e|ou} outros SN_1
(10) SN_1, e em particular SN_2
(11) Dentre SN2, SN_1, (esse padrão parece não se aplicar ao português)

(6) SN_1, particulièrement SN_2
(7) SN_1 comme Lista de SNs
(8) SN_1 tel Lista de SNs
(9) SN_2 {et lou} de autre SN_1
(10) SN_1 et notamment SN_2
(11) Chez le SN2, SN_1,

Esses padrões foram aplicados em um corpus constituído de resumos e títulos de artigos científicos produzidos por pesquisadores, engenheiros e técnicos das áreas de agricultura e indústria alimentícia, o corpus "[AGRO-ALIM]". O corpus possui 427.482 palavras, com uma média de 316 palavras por resumo (Jacquemin et al., 2002).

A avaliação dos pares extraídos mostrou uma alta qualidade das relações produzidas, com uma precisão de 82%, mas uma abrangência de 56%. A avaliação foi feita por padrão extraído, e a tabela 3 reproduz os resultados de alguns padrões semelhantes aos descritos em Hearst.

Padrão	Qtde de relações	Precisão
(5) SN_1 tais como Lista de SNs	210	86%
(7) SN_1 como Lista de SNs	90	69%
(8) SN_1 tais como Lista de SNs	36	90%
(9) SN_1 elou outros SNs	17	59%

Tabela 3: Resultados de alguns padrões de Morin e Jacquemin (2004)

Após a descoberta dos padrões, e conseqüente extração de pares semanticamente relacionados, Morin e Jacquemin (2004) apresentam uma técnica para a aquisição incremental das relações extraídas por meio da exploração de relações sintáticas, morfossintáticas e semânticas entre os termos extraídos. Embora interessante, o método se apóia em uma ferramenta altamente sofisticada chamada FASTR (Jacquemin, 1999), um parser transformacional para o qual não há equivalente na língua portuguesa.

O trabalho de Cederberg e Widdows (2003) consiste na utilização de modelos matemáticos (Latent Semantic Analysis – LSA) para medir a similaridade semântica entre as palavras.

Os autores realizam três experimentos: no primeiro deles, constroem um sistema extrator de hiperonímia que utiliza as 6 regras de Hearst (1998). A partir de uma amostra de 430.000 palavras do *British National Corpus,* são extraídas 513 relações, das quais 100 foram selecionadas para avaliação manual. Na avaliação, cada relação deveria ser pontuada de acordo com os seguintes critérios:

> 4. As relações estão corretas da maneira como foram extraídas.
> 3. As relações estão corretas após uma ligeira modificação, como mudança plural-singular ou a remoção de artigo.
> 2. As relações estão "potencialmente corretas" mas requerem um processamento difícil para a obtenção da relação correta. Por exemplo, o substantivo está correto mas há problemas no sintagma preposicional.
> 1. A relação está correta de alguma forma, mas é muito geral ou muito específica para ser útil.
> 0. A relação está incorreta.

Após a avaliação, 40% das relações foram pontuadas como 3 ou 4 (relações corretas). A fim de melhorar os resultados, Cederberg e Widdows aplicaram um filtro utilizando uma variante do método LSA[18]. Os novos resultados mostraram um aumento das relações classificadas como 3 ou 4 de 40% para 58%, o que sugere a efetividade do filtro.

Em uma tentativa de aumentar o número de relações identificadas, já que os padrões de Hearst são considerados pouco freqüentes nos textos, Cederberg e Widdows utilizaram um método já descrito em Widdows e Dorow (2002), que consiste em considerar a pista fornecida pela estrutura de coordenação da língua (elementos que aparecem em listas tendem a ser semanticamente similares) aliada a um método de agrupamento (*clusterização*). Para tanto, assume-se que, em uma frase como

(4) Este não é o caso de açúcar, mel, cravos ***e outras*** *especiarias que...*

que leva à identificação da relação

cravos < especiarias,

e em uma frase como

(5) *Navios carregados com noz-moscada ou canela,* ***cravos ou*** *coentro enfrentaram...*

[18] O método LSA (Latent Semantic Analysis) avalia em que medida as palavras *x* e *y* aparecem em contextos similares por meio da representação de palavras como pontos em um espaço vetorial. Palavras com significados relacionados devem ser representadas como pontos próximos.

como a relação entre *cravo* e *especiarias* já foi identificada, a hipótese de coordenação levaria a identificação das relações

noz-moscada<especiarias;
canela<especiarias;
coentro< especiarias

Na etapa final do trabalho, Cederberg e Widdows (2003) aplicaram novamente o filtro LSA nos resultados da extração, que incluem aqueles obtidos com a utilização da pista de coordenação. De 260 relações avaliadas, 166 (64%) foram consideradas corretas (pontuação 3 ou 4), o que mostra o sucesso na combinação das técnicas.

Snow, Jurafsky e Ng (2005), partindo da crítica de que os padrões léxico-sintáticos de Hearst (1998), embora amplamente utilizados em outros trabalhos, têm limitações quanto à abrangência (isto é, são poucos padrões) e quanto à forma de identificação (em geral, os padrões são identificados manualmente), propõem a utilização de aprendizagem de máquina para substituir este conhecimento construído manualmente. Em termos gerais, a abordagem de Snow et al.(2005) baseia-se em (a) coletar pares de substantivos, no corpus, que identifiquem relações de hiperonímia, utilizando a WordNet; (b) coletar, para cada par, frases em que ambos os substantivos apareçam; (c) realizar um *parsing* dessas frases para a extração automática de padrões; (d) treinar um classificador de hiperônimos utilizando esses resultados. Embora o trabalho pareça muito interessante, não temos como comparar os resultados de Snow et al. com os demais apresentados aqui, visto a forma de avaliação ser bastante diferente.

Em suma, embora diversos trabalhos venham propondo a identificação automática em textos de relações de hiperonímia, os padrões descritos originalmente em Hearst (1992, 1998) têm se mostrado os mais produtivos, sendo amplamente repetidos em combinação com outras técnicas.

A principal crítica à abordagem de Hearst é sua pouca abrangência, isto é, provavelmente nem todas as relações semânticas relevantes para uma ontologia são expressas por meio de pistas textuais – e talvez nem todas as relações de hiperonímia de um domínio. Por outro lado, a metodologia apresenta a grande

vantagem de oferecer grupos de palavras já rotulados com um hiperônimo, e não simplesmente aglomerados de palavras.

Trabalhos como os de Cederberg e Widdows (2003) e Snow et al. (2005) tentam conciliar os padrões com outras técnicas, a fim de aumentar a precisão e abrangência dos resultados, mas os dados, até o momento, sugerem que tais melhorias são pouco significativas.

Já o trabalho de Morin e Jacquemin (2004) apresenta um sistema capaz de extrair automaticamente do corpus padrões léxico-sintáticos para a expressão de relações semânticas. Para tanto, o sistema utiliza algoritmos e o cálculo estatístico de medida de similaridade. Porém, os padrões encontrados são muito semelhantes aos de Hearst, e resta saber se o processo extração automática de regras teria um desempenho tão eficaz na identificação outras relações semânticas, como a meronímia, que não têm sido tão exploradas.

As relações de hiperonímia identificadas por Morin e Jacquemin (2004) apresentam, em termos gerais, resultados bastante superiores aos de Hearst e de Cederberg e Widdows. Porém, uma comparação exata entre os trabalhos não é possível por diversas razões.

A primeira delas diz respeito ao tipo de avaliação realizada em cada trabalho. Hearst (1998) e Cederberg e Widdows (2003) avaliam a precisão das relações por meio de uma escala (parecida, mas não idêntica) de aceitação das relações identificadas, que vai do acerto total ao erro total; Morin e Jacquemin (2004) utilizam medidas de precisão e abrangência. Por outro lado, Hearst apresenta seus resultados por padrão léxico-sintático – especificamente, apresenta os resultados obtidos com apenas um padrão. Morin e Jacquemin também apresentam os resultados obtidos por padrão identificado, mas Cederberg e Widdows (2003) apresentam os resultados gerais, isto é, não sabemos o desempenho de cada regra.

O segundo obstáculo para uma comparação adequada diz respeito ao corpus: Hearst utiliza textos jornalísticos; Cederberg e Widdows (2003), uma amostra do *British National Corpus*, um corpus diversificado; e Morin e Jacquemin (2004) um corpus relativamente "controlado", composto por resumos de artigos técnicos, de um domínio específico.

Por fim, as diferenças quanto ao idioma também devem ser levadas em consideração: o trabalho de Morin e Jacquemin (2004) tem o francês como língua-alvo, e os trabalhos de Hearst e de Cederberg e Widdows voltam-se para o inglês.

5
Metodologia

Neste capítulo, apresento o corpus e os padrões léxico-sintáticos utilizados para a identificação de relações de hiperonímia.

5.1.
O corpus

Para a extração das relações semânticas, foi utilizado um corpus de 11 MB (1.846.502 palavras), composto por textos da área de saúde pública disponíveis na Internet. Os textos, de registro formal, pertencem a diferentes gêneros textuais: artigos acadêmicos, cartilhas, manuais, textos de divulgação, textos didáticos e jornalísticos. A opção pela heterogeneidade quanto ao gênero – isto é, a escolha de textos com diferentes graus de complexidade quanto ao tema – se deve à tentativa de capturar diferentes "níveis" de informação. Isto porque é possível supor, por exemplo, que textos especializados, como artigos acadêmicos, já assumem um conhecimento compartilhado de nível mais básico, de maneira que não precisam explicitar informações do tipo "*enzimas são substâncias*", mas sim "*colagenase é uma enzima*". Essas informações mais básicas, por sua vez, espera-se que sejam explicitadas em textos didáticos e / ou textos de divulgação. Desse modo, tendo em vista o objetivo final de elaboração de ontologia, o que se pretende com um corpus com essas características é que os diferentes níveis de conhecimento emerjam do texto, caracterizando as diferentes categorias da ontologia.

5.1.1.
O pré-processamento do corpus

Para a aplicação dos algoritmos de identificação de padrões sobre o corpus, é necessário que ele já tenha passado por uma série de etapas:

Etiquetagem morfossintática: é fundamental que o corpus contenha etiquetas de classes gramaticais (POS tags). Para isso, o corpus foi anotado pelo etiquetador automático do parser PALAVRAS (Bick, 2000).

Etiquetagem de Sintagmas Nominais: Já com as etiquetas de classes gramaticais, o corpus passou por um etiquetador automático de Sintagmas Nominais (Santos e Oliveira, 2005), já que as regras de identificação dos padrões são dependentes da segmentação em SNs.

Após o processo de etiquetagem automática, o corpus foi manualmente revisto, a fim de minimizar, principalmente, erros decorrentes da identificação / segmentação de nomes próprios.

5.2. Descrição dos padrões

O primeiro passo para a elaboração de uma ontologia é a indicação das relações semânticas desejadas. A princípio, foram escolhidas as relações hiperonímia/hiponímia, por possibilitarem a realização de inferências, e relações de co-referência, por oferecerem um tipo de definição, ainda que informal. A etapa seguinte é a identificação, no texto, de padrões léxico-sintáticos que expressam essas relações semânticas. Nessa etapa, a aquisição da informação é semi-automática, pois precisa da avaliação do pesquisador sobre o corpus para a identificação dos padrões relevantes. Em um momento posterior, quando os padrões já estão identificados, é possível utilizar mecanismos de identificação e extração automáticas.

Como visto na seção 4.2.1, Hearst (1998) apresenta seis pistas textuais para a extração da relação de hiperonímia:

(i) NP_0 such as NP_1 {, NP_2 ... , (and | or) NP_i}
(ii) such NP_0 as {NP ,}* {(and | or)} NP
(iii) NP {, NP}* {,} or other NP_0
(iv) NP {, NP}* {,} and other NP_0
(v) NP_0 {,} including { NP ,}* {or | and} NP
(vi) NP_0 {,} especially { NP ,}* {or | and} NP

Neste trabalho, utilizei três pistas de Hearst – pistas (i), (iii) e (iv) –, com algumas modificações, e descartei as demais por serem pouco produtivas. Além disso identifiquei, por meio da observação do corpus, mais três outros padrões:

tipos de SN_0: SN_1 { , SN_2 ... ,} (e | ou) NP_i ;

SN_0 ***chamado/s/a/as*** SN_1;

SN ***conhecido/s/a/as como*** SN.

As seções seguintes detalham cada um dos padrões utilizados.

5.2.1. O padrão "tais como"

O padrão (i) de Hearst (1998) – "*such as*" –, pode ser literalmente traduzido para "*tais como*". Porém, na língua portuguesa, freqüentemente apenas o "*como*" é utilizado neste tipo de construção, como ilustram (1) e (2):

(1) *A tentativa posterior de clonar outros mamíferos* ***tais como*** *camundongos, porcos, bezerros*,....

(2) *A tentativa posterior de clonar outros mamíferos* ***como*** *camundongos, porcos, bezerros*,....

Ou seja, para que o padrão revele uma quantidade significativa de relações de hiperonímia no português, é preciso considerar a variante "*como*". Porém, se há um ganho do ponto de vista da abrangência, uma vez que mais relações podem ser identificadas, do ponto de vista da precisão essa inclusão é um complicador: "*como*" é uma palavra que se enquadra em diferentes classes gramaticais, dificultando o trabalho dos etiquetadores automáticos e, conseqüentemente, acarretando problemas na identificação do padrão desejado.

Pela gramática tradicional, "*como*" pode ser advérbio, preposição acidental, pronome relativo ou conjunção. Quando conjunção, pode ser subordinativa – adverbial ou integrante – ou coordenativa.

O quadro 2 ilustra cada um dos casos, com a respectiva etiqueta morfossintática atribuída pelo conjunto de etiquetas do parser PALAVRAS e pelos etiquetadores do projeto Lácio-Web:

Frase	classe grammatical	PALAVRAS	Lácio-Web
...não sabiam **como** se proteger...	Conj. Sub. Integr.	ADV	ADV-KS
Como é muito difícil comprovar...	Conj. Sub. Adv.	KS	KS
A expectativa tanto em países desenvolvidos **como** em países em desenvolvimento...	Conj. Coord.	<parkc-2> DV Tanto como (par)	KC
... a doença periodontal têm **como** conseqüência o edentulismo...	Advérbio	ADV	ADV

...cabe aqui uma outra frase **como** resumo do pensamento de...	Prep. acidental	ADV	PREP
... verdade no modo **como** ele interpreta aquela dualidade...	Pron. Relativo	ADV	PRO-KS

Quadro 2: Exemplos de etiquetas atribuídas ao "*como"* por etiquetadores automáticos

Porém, o *como* que nos interessa não se encontra em nenhum dos casos exemplificados. Aliás, ele quase não aparece nas gramáticas. Não por acaso, ele também não recebe nenhuma etiqueta especial pelos etiquetadores automáticos. Na frase (3)

(3) *Com a entrada de [instrumentos] **como** [flauta], [bandolim] e [cavaquinho], estava completa a gestação do chorinho.*

o "*como*" foi etiquetado como preposição (PREP) pelos etiquetadores Brill e TreeTagger[19] e como <rel> <ks> <prp> ADV pelo Palavras – a mesma etiqueta atribuída aos termos em negrito nos exemplos abaixo[20]:

*(a) ...repasse e armazenamento de dados, **conforme** descrição**...***
*(b) Você não encara aniversários **como** mais um ano de vida*
*(c) Esta base de dados não tem **como** proveniência a Lista Telefônica..*
*(d) é artista na forma **como** agrada ao seu amante.*
(e) *resposta ao que interpretei **como** um apelo de Deus.*

O "como" do exemplo (3), aquele que nos interessa na identificação da relação de hiperonímia, pode ser utilizado no lugar (ou acrescido de) "por exemplo":

"*Com a entrada de instrumentos **como por exemplo** flauta, bandolim...*

Neste caso, trata-se de um *como* que pode ser classificado como uma "palavra denotativa", do mesmo modo que seria a expressão "por exemplo"[21]. Ou

[19] Os etiquetadores estão disponíveis no sítio do projeto Lácio-Web: http://nilc.icmc.sc.usp.br/lacioweb/

[20] É importante destacar que a igualdade entre as etiquetas do PALAVRAS só acontece porque, durante a utilização online do sistema, foi selecionada a opção "*morphological tagging*" Quando se escolhe a opção "full morphosyntactic parse", os diferentes "*como*" dos exemplos são desambiguizados, e o "*como*" da frase (3) recebe a etiqueta a etiqueta ADV @AS-N<, que é interpretada como uma construção elíptica; uma oração adverbial em que o verbo ser está elíptico: "*instrumentos como* [o são] *flauta, bandolim...*"

[21] Pereira (1995) aponta para a polêmica suscitada pela classe das denotativas, que ora são colocadas à parte, ora incluídas entre os advérbios, e ora não são sequer mencionadas. Concordamos com Pereira quanto à necessidade de classificação à parte das denotativas, uma opção coerente uma vez que há, na língua, diversas palavras cuja classificação pode variar conforme o emprego. Palavras denotativas são um recurso que a língua oferece, e por isso devem ter status próprio, sendo desnecessário o estabelecimento de uma classificação granular do tipo "*denotativa de...*" (cf. Oliveira e Freitas, 2006).

seja, o *como* palavra denotativa, semelhante a "tais como" e equivalente a "por exemplo", tem chances mínimas (senão nulas) de receber uma etiqueta PDEN – palavra denotativa (etiqueta inexistente no parser PALAVRAS mas disponível no conjunto de etiquetas do projeto Lácio-Web).

Conseqüentemente, uma busca pelo padrão "*SN como SN*", que considera a etiqueta PDEN de "*como*", provavelmente leva a um alto índice de precisão – e, do mesmo modo, a desconsideração da etiqueta leva a inúmeros erros.

Uma pista já utilizada por Hearst (1998) para a identificação do "*tais como*" – no caso do inglês – é a presença de coordenação (lista de SNs) após o "*tais como*". Nos exemplos anteriores, de fato, o único caso em que há ocorrência de lista após o "*como*" é justamente o caso que nos interessa. Porém, embora a coordenação seja pista eficaz e prática, pois elimina a dependência de um etiquetador altamente preciso, ela não é suficiente.

Nos exemplos (4) e (5) há uma seqüência de "SN como {lista de SN}" que não corresponde ao padrão desejado:

> *(4) O uso da bebida compromete a vida física e moral do alcoólico, representada pela perda de suas qualidades morais e de suas [responsabilidades] **como** [pai], [esposo] e [trabalhador].*

> (5) *O modelo central foi considerado satisfatório quando os resíduos não apresentaram mais associação com as variáveis meteorológicas e a série de resíduos em função de o tempo não evidenciou mais nenhum [padrão] **como** [tendência], [sazonalidade] ou [autocorrelação].*

Além disso, embora pouco freqüentes, as estruturas em que o "*como*" é palavra denotativa, mas vem seguido por um único SN – e não por uma lista –, também deixam de ser identificadas quando se considera exclusivamente a pista da coordenação, como mostram (6) e (7):

> *(6) A falta de [minerais] **como** [o ferro] pode causar uma anemia.*

> *(7) ... o que torna ainda mais importantes [iniciativas] **como** [a Campanha de Carnaval 2003], que buscam estimular...*

A inclusão do padrão "como_PDEN" nos deixa com um problema: por um lado, é altamente confiável como expressão de relação de hiperonímia e muito

---

Considerando o objetivo primeiro de identificação automática deste tipo de "*como*", não faz diferença se ele é visto como um advérbio que introduz oração elíptica ou como uma palavra denotativa – o que importa é que receba uma etiqueta que o diferencie dos demais "*como*".

mais freqüente na língua do que o padrão "*tais como*" (o corpus de saúde utilizado contém cerca de 2700 ocorrências de "como_PDEN" contra apenas 232 ocorrências de "tais como"); por outro lado, o sucesso de sua identificação depende de um fator externo – depende de um etiquetador capaz de reconhecer o "como_PDEN" ou o "*como*" que introduz uma oração adverbial elíptica. Devido ao grande número de ocorrências como_PDEN (mais de dez vezes o número de ocorrências de "tais como"), decidimos re-etiquetar, manualmente, todos os "*como*" que fossem palavra denotativa.

Deste modo, para o padrão original

$$NP_0 \text{ such as } NP_1 \{, NP_2 \ldots, (\text{and} \mid \text{or}) NP_i\}$$

utilizamos, para o português,

$$\text{(I) } SN_0 \text{ (tais como} \mid \text{como\_PDEN) } SN_1 \{ , SN_2 \ldots ,\} \text{ (e} \mid \text{ou) } SN_i$$

capaz de extrair relações de estruturas como

(8) *...e [distúrbios metabólicos], **como_PDEN** [hiponatremia], [hipoglicemia] e [hipocalcemia], pois a infecção ...*

(9) *O estágio adulto é mais específico de [grandes mamíferos] **como_PDEN** [eqüinos], [antas] e [capivaras] e, eventualmente, ...*

mas incapaz de extrair informação de (10), (11) e (12)

(10) *... pode pensar na vacina **como_ADV** uma pequena armadilha: ao mudar de forma, o vírus...*

(11) *... estendia-se pela capital **como_ADV** uma densa rede ...*

(12) *... fica evidente o modo **como_ADV** os usuários tornam-se ...*

Além da especificidade do como_PDEN, o padrão "como/tais como" (mas não apenas ele, como será visto mais tarde) apresenta outro fator complicador, já notado por Hearst (1998): a ambigüidade de estruturas que contêm sintagmas preposicionados (SPrep). Em estruturas como

(13) *Incorpore à sua rotina [atividades redutoras de o estresse], **como** [exercícios], [ioga], [meditação],[jardinagem] ...*

*(14) **[**Infecções por bactérias**]** **como** **[**a Salmonella**]** e **[**a Shighella**]** ...*

*(15) O tratamento é feito por meio de **[**a administração de medicamentos**]** **como** **[**o oxamniquine**]** e **[**o praziquantel**]**, porém, a melhor maneira de enfrentar...*

Pela regra (I), seriam extraídas, respectivamente, as relações

(13 a) exercícios < atividades redutoras de o estresse
(13 b) ioga < atividades redutoras de o estresse
(13 c) meditação < atividades redutoras de o estresse
(13 d) jardinagem < atividades redutoras de o estresse
(14 a) Salmonella < infecções por bactérias
(14 b) Shighella< infecções por bactérias
(15 a) oxamniquine < a administração de medicamentos
(15 b) praziquantel < a administração de medicamentos

em que apenas as relações extraídas da frase (13) estão corretas. A solução foi criar, ao lado do SN hiperônimo (SN Hiper), o SN HHiper, que considera SN hiperônimo o primeiro N à esquerda do "*como / tais como*". Com essa alteração, as relações extraídas de (14) e (15) ficam corretas

(14 a') Salmonella < bactérias
(14 b') Shighella< bactérias
(15 a') oxamniquine < medicamentos
(15 b') praziquantel < medicamentos

mas, por outro lado, as relações de (13) se tornam erradas:

(13 a') exercícios < estresse
(13 b') ioga < estresse
(13 c') meditação < estresse

A análise do corpus mostrou, porém, que uma outra alteração na regra permitiria ainda mais acertos na identificação das relações de hiperonímia: quando houver vírgula antecedendo o "*como / tais como*", o hiperônimo considerado é o

SN Hiper "tradicional", isto é, o SN completo, e não apenas o primeiro substantivo à esquerda de "como / tais como", como ilustram os exemplos (16) e (17).

(16) *… procurou- se obter [outros dados relativos à sífilis materna]* ***,*** ***como*** *[a titulação do VDRL no parto] , em a tentativa de …*

(17) *…poderiam se correlacionar com [os cuidados em o período não-reprodutivo] ,* ***como*** *[o uso da TRH] .*

De fato, parece haver uma motivação discursiva para essa diferenciação: a vírgula empregada após os sintagmas hiperônimos formados por mais de um substantivo indicaria uma pausa necessária para a retomada de toda a informação veiculada no sintagma anterior que, por sua vez, estará relacionada ao SN hipônimo. Já nos casos de SN hiperônimos com mais de um substantivo, mas que não aparecem seguidos de vírgula, os SNs hipônimos estariam relacionados apenas ao último N do sintagma, o N mais próximo, como ilustra o exemplo (18):

(18) *... e ocorre o funcionamento inadequado dos [órgãos vitais]* ***como*** *[fígado] e [rins].*

A regra final utilizada na identificação do padrão "como/tais como" foi, portanto, desmembrada em duas:

(Ia) SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) SNi

(Ib) SN Hiper, (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) SNi

5.2.2. O padrão "e/ou outros"

A identificação das relações expressas pelo padrão "e outros", tratado em Hearst (1998) por meio das pistas (iii) e (iv), também sofre com problemas decorrentes da ambigüidade do sintagma preposicionado, como ilustram (19-22):

(19) *... [a evolução de referenciais teóricos postos à disposição de [educadores]]* ***e outros*** *[pesquisadores].*

(20) *... [o acesso a [serviços de [laboratório]]]* ***e outros*** *[meios diagnósticos]*

(21) ... [a experiência subjetiva com [o LSD-25]] ***e outros*** *[alucinógenos]*

(22) *... pode contribuir para [a maior ocorrência de [doenças cardiovasculares]], [cânceres]* ***e outras*** *[enfermidades] ...*

Neste caso, porém, a dificuldade de segmentação não está no SN hiperônimo, mas nos SNs hipônimos. A solução que encontramos para minimizar esse problema foi criar, ao lado do SN HHiper, o SN HHipo: é considerado SN hipônimo o primeiro N anterior à expressão "e/ou outros" e, no caso de uma coordenação de hipônimos, a estrutura HHipo se aplicará sempre ao sintagma mais à esquerda da relação. Nos exemplos (19-22) seriam extraídas, portanto, as relações:

(19') educadores < pesquisadores
(20') * laboratório < meios diagnósticos
(21') LSD-25 < alucinógenos
(22 a') doenças cardiovasculares < enfermidades
(22 b') cânceres < enfermidades

Como é possível perceber, nem sempre a estratégia HHipo obterá sucesso – como é o caso da relação (20') – , já que as estruturas são de fato ambíguas e, freqüentemente, o nosso conhecimento de mundo será o responsável pela segmentação correta do sintagma. Porém, ainda que existam erros, a estratégia é capaz de eliminar grande parte deles, o que não aconteceria se utilizássemos os sintagmas hiperônimos / hipônimos tradicionais, como faz Hearst (1998). Desse modo, para a identificação do padrão "e/ou outros", substituímos as regras originais, em inglês, (iii) e (iv), por:

(II) SN HHipo { ,SN Hipoi } * { , } elou outros SN Hiper

Porém, diferentemente do padrão "como/tais como", o padrão "e/ou outros" apresenta uma peculiaridade semântica/discursiva: algumas vezes, o sintagma candidato a hiperônimo está relacionado a um termo elíptico, ausente na coordenação mas presente em outra oração (23) ou mesmo em outro parágrafo

(24). Nestes casos, o SN após "e/ou outros" não se comporta como um hiperônimo, mas como um termo anafórico que retoma um outro termo que tanto pode ser seu hipônimo, um equivalente do termo referido ou uma repetição do próprio termo, numa estratégia coesiva:

(23) *... nunca se deve esquecer que ao drogado restam, como amigos e companheiros, apenas os [traficantes]* ***ou outros [*** *viciados].*

(24) *Da mesma forma que para a LV canina, o sacrifício do* ***cão*** *positivo (…) também é recomendado por não existir tratamento eficaz e o animal também constituir importante reservatório dessas doenças para o ser humano. // (...), foram detectados 2.003 animais falsos negativos e que, assim, não foram sacrificados. Não se pode deixar de considerar que a permanência desses animais no ambiente epidêmico pode certamente ter comprometido a eficácia (…), contribuindo para a manutenção de focos da doença e, conseqüentemente, fontes de infecção para [pessoas]* ***e outros*** *[cães].*

Embora pouco freqüentes, os erros decorrentes dessa estratégia coesiva indicam que, no padrão "e/ou outros", a expressão da relação de hiperonímia não é tão garantida quanto no padrão "como/tais como".

5.2.3. O padrão "tipos de"

A partir da observação do corpus, percebemos que o padrão "tipos de" também expressa relação de hiperonímia:

(25) Existem dois ***tipos de*** *[cromossomos gigantes]: [cromossomos politênicos] e[cromossomos plumulados].*

(26) No sangue se medem essencialmente três ***tipos de*** *[colesterol]: [o colesterol total], [o colesterol HDL] e [o colesterol LDL].*

Porém, diferentemente dos anteriores, o padrão "tipos de" não apresenta problemas de ambigüidade relativos ao sintagma preposicionado, nem particularidades de natureza discursiva ou coesiva – o que significa que as relações identificadas são altamente confiáveis. A regra correspondente ao padrão é

(III) tipos de SN Hiper: SN_1 { , SN_2 ... ,} (e | ou) SNi

5.2.4.
O padrão "chamado/a/os/as"

Este padrão também foi descoberto a partir da observação do corpus:

(27) ... e nele existe uma [substância] ***chamada*** *[benzopireno].*

(28) Este fato tem sido descrito com freqüência na [doença mental] ***chamada*** *[esquizofrenia].*

Nele, também há dificuldade na identificação da relação decorrente da ambigüidade do sintagma preposicionado, e, novamente, foi utilizada a estrutura HHiper (regra IV):

(IV) SN HHiper chamado/s/a/as (de) SN Hipo

5.2.5.
O padrão "conhecido/a/os/as como"

Foi investigado ainda o padrão "conhecido como". Neste caso, porém, o objetivo não é a expressão de hiperonímia, mas de co-referência entre os termos. Isto é, buscamos aqui obter sinônimos, ou até mesmo definições, para os termos envolvidos nas estruturas, como mostram (29) e (30):

(29) Cerca de 95% dos adultos já tiveram a virose mononucleose infecciosa ou [angina monocítica], também ***conhecida como*** *[doença do beijo].*

(30) ..., protege contra [o tétano neonataltétano], ***conhecido como*** *[mal dos sete dias].*

Com este padrão as relações extraídas são de co-referência, e têm a forma

(29') angina = doença do beijo

(30') tétano neonataltétano = mal dos sete dias

Para a identificação automática desta estrutura, a regra utilizada foi

(V) SN Hiper *conhecido/s/a/as como* SN Hipo.

Neste capítulo apresentei os padrões utilizados na extração de relações semânticas do corpus, que irão organizar a ontologia de domínio. Para tanto, utilizei três padrões apresentados originalmente em Hearst (1992), introduzindo algumas alterações:

- inclusão de um sintagma hiperônimo SN HHiper para casos em que o SN contém mais de um substantivo;
- acréscimo da estrutura "como_PDEN" ao lado da regra original "tais como"
- alternância entre a utilização de SN HHiper e SN Hiper na regra "como/tais como" em função do emprego da vírgula.

Além disso, a partir da observação do corpus, acrescentei mais três padrões: dois para a identificação de hiperonímia – "tipos de" e "chamado/a/os/as" – e um para a identificação de co-referência – "conhecido/a/os/as como".

As regras para a identificação das relações têm a seguinte estrutura:

(Ia) SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) SNi

(Ib) SN Hiper, (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) SNi

(II) SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper

(III) tipos de SN Hiper: SN_1 { , SN_2 ... ,} (e | ou) SNi

(IV) SN HHiper chamado/s/a/as (de) SN Hipo

(V) SN Hiper *conhecido/s/a/as como* SN Hipo

6
Resultados

A análise dos resultados foi realizada em 3 etapas. Na 1ª etapa, o objetivo principal foi identificar os erros de natureza sintática, sem preocupação com a utilidade / exatidão das relações extraídas. Isto é, nesta etapa, parto do pressuposto de que os padrões investigados expressam, de fato, relações de hiperonímia e de co-referência, ainda que não sejam relações "convencionais" de um ponto de vista lexicográfico. Desse modo, foram consideradas corretas relações como

sensibilidade<condição

reforma de um jardim<trabalhos voluntários

Foram considerados erros casos em que:

(a) a relação extraída não estava correta devido à ambigüidade do sintagma preposicionado. No exemplo (1), a relação extraída é *transmissão de HBV<patógeno*, e não *HBV<patógeno*:

(1) ... **[***transmissão de o HBV vivo***]** ***e outros [****patógenos****]*** ...

(b) uma estrutura adverbial deslocada da ordem direta (encaixada) assume a forma do padrão buscado. No exemplo (2) são extraídas as relações *países de prevalência relativamente baixa* > *China, taxas* e no exemplo (3) as relações *ingestão* > *leptospirose, hepatite A, hepatite E*

*(2) agora, mesmo em **[**países de prevalência relativamente baixa**]** **como** a **[**China**]**, **[**as taxas**]** em algumas cidades chegam a quase 20%.*

*(3) as inundações aumentam os riscos de aquisição de doenças infecciosas transmitidas por água contaminada, através de contato ou **[**ingestão**]**, **como [**leptospirose**]**,**[**hepatite A**]**, **[**hepatite E**]**, ...*

(c) elipse de algum termo. No exemplo (4), são extraídas *amplo número de indivíduos* > *grupos comunitários e de trabalhadores, estudantes, grupos étnicos isolados, centros religiosos*

(4) *... e atender a um* **[***amplo número de indivíduos***]** *,* ***como*** **[***grupos comunitários e de trabalhadores***]** *,***[***estudantes***]** *,* **[***grupos étnicos isolados***]** *ou* **[***centros religiosos***]**.

(d) presença de uma oração no interior do sintagma hiperônimo ou hipônimo. No exemplo (5), são extraídas *concentração energética mínima* > *sopas, mingaus*.

(5) *...preparações que não atinjam* **[***esta concentração energética mínima***]***,* ***tais como*** **[***sopas***]** *e***[***mingaus***]**

Ou seja, nessa etapa, foram considerados erros os padrões extraídos que correspondem a uma estrutura sintática diferente da estrutura alvo ou em que peculiaridades sintáticas contribuem para um desvio do padrão-alvo, já que, em termos semânticos, assumimos que os padrões expressam as relações desejadas, ainda que de uma forma pouco convencional – como dito antes, assumo que explicações serão sempre parciais.

Com esses critérios, foi feita uma avaliação manual dos resultados (tabela 4):

Padrão	Quantidade de Relações	Acertos
como/tais como	2428	1824 (75%)
e outros	394	321 (81.4%)
tipos de	21	18 (85%)
chamado	89	81 (91%)
conhecido como	76	38 (50%)
TOTAL	**3008**	**2282 (75.8%)**

Tabela 4: Resultados das extrações por padrão

6.1. Análise dos erros sintáticos

A aplicação em separado de cada regra mostrou que uma análise dos erros que também considerasse cada padrão isoladamente seria vantajosa, tendo em vista uma futura eliminação dos erros mais previsíveis. Em comum, todas as

estruturas apresentaram erros decorrentes da ambigüidade sintática na junção do SPrep – erro chamado de HHiper – , mas houve diferenças interessantes.

Com relação ao padrão "como/ tais como" (tabela 5), mais da metade dos erros foi decorrente da ambigüidade do SPprep, conforme já previsto. A surpresa foi o número relativamente alto (29%) de erros resultantes da presença de uma oração no SN Hiper. Como o modelo de SN utilizado na identificação dos padrões não comporta orações (Freitas et al., 2005), tais erros jamais seriam eliminados com a metodologia empregada[22]. Por outro lado, a utilização de um modelo de SN que levasse em consideração orações aumentaria de maneira considerável a precisão dos resultados.

Tipo de erro/ padrão "como/ tais como"	**Frase-exemplo**	**Relação extraída**	**Qtde de erros**
Or. encaixada	alimentos especiais são dados à [criança doente], **tais como** [chás] , [água de coco] e [sopas ralas]	criança doente > chás, água de coco, sopas	**175 (29%)**
Erros HHiper	a ocorrência de sintomas de [abstinência], **como** [náusea] , [suor] , [tremores] e [ansiedade]	abstinência > náusea, suor, tremores, ansiedade	**370 (61.5%)**
Outros erros	facilita o aparecimento de [doenças respiratórias] como [pneumonias] e [diarréias]	doenças respiratórias > pneumonias, diarréias	**56 (9.3%)**
Total	**--**	**--**	**601**

Tabela 5: Análise dos erros sintáticos do padrão "como/tais como"

Já no padrão "e/ou outros" os diferentes tipos de erros estão distribuídos de maneira relativamente homogênea. Diferentemente do que aconteceu no padrão "como/tais como", erros sintáticos – como o erro HHiper – aparecem com a mesma freqüência de erros de natureza semântica-discursiva. Isto é, diferentemente do "como/tais como", a identificação do padrão "e/ou outros", por si só, não garante a extração de uma relação de hiperonímia, como mostra a tabela 6. Quase um terço dos erros é decorrente de uma estratégia discursiva na qual, dada uma lista de elementos coordenados, o elemento hiperônimo posterior a "e outros" não faz referência a toda a lista, mas apenas ao(s) último(s) elemento(s) da lista. Em seguida aparecem erros decorrentes de uma anáfora que retoma como

[22] O modelo de SN descrito em Freitas et al. (2005), chamado SN lexical, tem como objetivo gerar termos indexadores para sistemas de recuperação de informação e, por isso,

hiperônimo um termo que não o é. Nos erros "outros" encontram-se principalmente relações decorrentes da presença de um adjunto adverbial anterior ao início da coordenação, cuja estrutura se confunde com a da lista.

Tipo de erro/ padrão "e/ou outros"	Frase-exemplo	Relação extraída	Qtde de erros
anáfora	A maioria das *mães* identificou aspectos positivos e benéficos dos projetos, como (...) compartilhar a aprendizagem com o marido, família, amigos **e outras mães** de bebês prematuros.	mães de bebês prematuros > marido, família, amigos	**17 (25%)**
Hiperônimo é o último substantivo da coordenação	...calhas, caixas d'água, bromélias **e outros** vegetais que acumulam água....	vegetais > calhas, caixas d'água, bromélias	**20 (29%)**
Erros HHiper	a instituição de um programa de controle de a anemia falciforme **e outras** iniciativas governamentais têm sido	anemia falciforme > iniciativas governamentais	**20 (29%)**
Outros erros	Em setembro, a British American Tobacco, a Philip Morris, a Japan Tobacco **e outras** companhias lançaram...	companhias > setembro, British American Tobacc, Japan Tobacco	**11 (16%)**
Total	**--**	**--**	**68**

Tabela 6: Análise dos erros sintáticos do padrão "e/ou outros"

Já o padrão "tipos de", embora pouco produtivo – apenas 21 ocorrências – apresentou um altíssimo grau de precisão. Os três únicos erros resultam de uma elipse do núcleo nominal, como pode ser observado no quadro 3.

Frase	Relações extraídas
...estudos iniciais com três **tipos de** [tumor]: [cérebro], [cólon[e [cabeça] e pescoço	cérebro<tumor cólon< tumor cabeça<tumor

Quadro 3: Erros obtidos com o padrão "tipos de"

Na verdade, como esses erros aparecem na mesma frase, poderiam ser considerados um único erro, ao invés de três. Além disso, é importante ressaltar que a baixa ocorrência desse padrão se deve, em grande parte, à estrutura do SN identificado. Em frases como (6)

---

caracteriza-se por ser uma mínima unidade lingüística com alto poder discriminatório, cujo núcleo deve ser uma única palavra lexical.

(6) *Existem três grandes* ***tipos de*** *conjuntivite: alérgica, infecciosa e aquela desencadeada por fatores externos.*

os hipônimos "(conjuntivite) alérgica" e "(conjuntivite) infecciosa" não são recuperados porque não contêm o núcleo nominal "conjuntivite", que está elíptico. Como este tipo de construção não parece ser incomum na língua, é possível que muitas relações não tenham sido identificadas.

O padrão "chamado" também obteve um alto percentual de acertos, e os poucos erros foram todos decorrentes da ambigüidade da estrutura com SPrep (tabela 7).

Tipo de erro/ padrão "chamado"	**Frase-exemplo**	**Relação extraída**	**Qtde de erros**
Erros HHiper	seqüenciaram duas regiões de um importante gene de o vírus de a Aids **chamado** de POL	POL< AIDS	**8 (100%)**
Total	**--**	**--**	**8**

Tabela 7: Erros obtidos com o padrão "chamado"

Por fim, a grande maioria dos erros do padrão "conhecido/a/os/as" (81%) foi decorrente do tipo de relação extraída. Lembro que, com este padrão, o objetivo não é a identificação de relações de hiperonímia, mas de co-referência. Contudo, o grande número de erros indica que o padrão é bastante ambíguo na identificação deste tipo de relação semântica: ora representa co-referência (7), ora representa hiperonímia (8)

(7) *... ou em* **[***vesículas esféricas de gordura***]** *, conhecidas como* **[***lipossomas***]** *, empregadas por serem compatíveis com o organismo ...*

(8) *aplicar em o tórax de o paciente um choque elétrico com* **[***um aparelho***]** *conhecido como* **[***desfibrilador***]**.

Devido ao baixo índice de acerto, o padrão "conhecido como" foi excluído da metodologia, o que nos deixou com um índice total de acertos de 76.4%.

Em termos gerais, a primeira etapa da análise dos erros evidenciou que a eliminação da ambigüidade do SPrep é de grande valia para um aumento na

precisão dos resultados, já que este é um tipo de erro presente em duas das estruturas investigadas. Além disso, um modelo de SN que considere orações encaixadas também levaria a um aumento na precisão. Do ponto de vista semântico-discursivo, a análise dos erros do padrão "e/ou outros" sugere que uma das formas de se aumentar a precisão seria considerar apenas o último elemento da lista de coordenação como hipônimo, e não todos os elementos da lista – e com isso eliminaríamos cerca de 30% dos erros. Esta solução pode ser interessante se integrada a um sistema maior, que utilize outros tipos de informação. Neste trabalho, como as regras são a única fonte de informação, perderíamos muito em recuperação, pois uma série de relações corretas deixariam de ser identificadas. Por isso, a regra "e/ou outros" foi mantida sem alterações.

Embora coerente com o ponto de vista teórico assumido, o critério de erro utilizado é pouco útil em dois aspectos importantes:

a) comparação de resultados: não há como comparar estes resultados com os apresentados em outros trabalhos (Hearst 1998; Widdows e Dorow 2003; Snow et al. 2005), devido à subjetividade da avaliação;

b) avaliação da funcionalidade: uma relação como

doença<fator ,

embora correta, é pouco significativa na elaboração de uma taxonomia e pode ser eliminada sem prejuízo (ou com um prejuízo mínimo) de informação.

6.2. Validação humana

A segunda etapa da avaliação teve como objetivo tornar os resultados "mais comparáveis" e "mais significativos": avaliadores[23] fizeram a validação de uma amostra dos resultados considerados "corretos" do ponto de vista sintático.

Das 2244 relações corretamente extraídas – assumindo o critério puramente sintático e excluindo os resultados do padrão "conhecido como" –,

[23] Participaram desta etapa 3 avaliadores, com formação em biologia, educação física e direito. A avaliação foi feita em conjunto, isto é, para cada relação avaliada, a resposta foi decorrente de um consenso entre os três.

uma amostra de 436 relações (cerca de 1/3) foi selecionada para avaliação humana. Numa pequena adaptação dos processos de validação utilizados por Hearst (1998) e Cederberg e Widdows (2003), foi pedido aos avaliadores que pontuassem as relações obedecendo aos seguintes critérios:

3	a relação está correta da forma como foi extraída
2	a relação está "um pouco" correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos, etc que o acompanham deixam a relação estranha.
1	a relação está correta em termos gerais; isto é, é muito geral ou muito específica para ser útil
0	a relação está errada

Porém, esses critérios, se, por um lado, pretendem oferecer alguma objetividade à tarefa de avaliação, por outro, não têm como assegurar a objetividade pretendida. No trabalho de Hearst, como a meta final é a inserção das categorias/relações na WordNet, a avaliação é relativamente mais simples, porque já existe um "padrão WordNet" de definição a ser seguido. No nosso caso, porém, freqüentemente é difícil distinguir entre uma "*relação correta*" (classificação 3) e uma relação"*muito específica para ser útil*" (classificação 1). De fato, grande parte da dificuldade da tarefa está justamente em determinar o que é o "ser útil". Relações como (a) e (b), abaixo, estão corretas ou são muito específicas – e pouco úteis?

(a) Superposição de tarefas<características da organização do trabalho
(b) Reavaliação do uso de anti-retrovirais<formas de recaptação do paciente

Além disso, no momento da validação, freqüentemente o senso comum difere do conhecimento enciclopédico, e então há divergências entre os avaliadores.

Por exemplo, do ponto de vista do senso comum, *cereais* podem ser um grupo alimentar; porém, do ponto de vista do conhecimento científico, *fibras* são um grupo alimentar, e não cereais. Qual deve ser o critério? A instrução dada aos avaliadores para que determinada relação fosse considerada correta é que a relação fosse verdadeira em algum mundo possível, isto é, existe pelo menos uma

circunstância em que a relação pode ser verdadeira. Com isso, *cereais* foi aceito como *grupo alimentar*. Os resultados da avaliação humana estão na tabela 8.

Classificação	Qtd de relações	Exemplos
3	320 (73.4%)	superóxido dismutase<enzimas suco<bebidas
2	15 (3.4%)	sofrimento<sentimentos inerentes à condição psicólogos<agentes da equipe
1	70 (16%)	proteção<valores queima de neurônios<comprometimentos
0	31 (7.1%)	setor público<serviços soco<traumas

Tabela 8: Resultados da avaliação humana

6.2.1. Filtro 1: substantivos gerais

Os resultados da avaliação indicam que a maioria das relações (73.4%) foi considerada correta da maneira como foi extraída, o que é um resultado muito bom. A maior parte dos erros está na categoria 1, e é decorrência de definições gerais demais ou específicas demais – e, conseqüentemente, pouco úteis. Neste caso estão relações cujo hiperônimo é um substantivo do tipo "fator", "termo" "elemento", "questão", aspecto", etc. Tais hiperônimos se enquadram na lista dos substantivos de sentido geral descritos em Marques (1995), e de substantivos-suporte descritos em Oliveira (2006): trata-se de substantivos com um alto grau de generalidade ou falta de especificidade, independentes de contexto temático.

De modo a eliminar tais relações gerais demais e pouco informativas, foi aplicado o 1º filtro, que elimina as relações cujo hiperônimo é um substantivo geral ou suporte.

Porém, alguns cuidados são necessários nesta etapa, pois os substantivos suporte descritos em Oliveira (2006) exercem a função de suporte justamente quando associados a complementos, que carregarão grande parte do significado do sintagma, deixando, conseqüentemente, o conteúdo do substantivo-suporte enfraquecido. Neste trabalho – nas relações extraídas dos corpus – , quando os substantivos-suporte estiverem acompanhados de complemento, eles serão mantidos, pois será justamente a presença do complemento a responsável por não deixar a relação extraída "vaga demais". No exemplo (c), a relação é eliminada,

pois é muito pouco informativa. Já a relação (d) é mantida, pois o adjetivo carrega a especificação necessária para que a relação seja considerada útil.

(c) osteoporose < fatores

(d) umidade < fatores climáticos

O complicador está no fato de que os substantivos-suporte são assim caracterizados justamente porque estão na presença de um complemento; isto é, quando utilizados sem complemento, podem funcionar como substantivos plenos em algumas situações. Porém, como assinala Oliveira (2006), tais situações são as de linguagens especializadas, jargões.

O problema passou a ser como identificar se a palavra candidata a substantivo-suporte / genérico estava de fato sendo empregada como tal ou se funcionava como substantivo pleno. Uma solução simples, embora não automática, foi simplesmente assumir que os substantivos-suporte só serão plenos quando usados em domínios específicos – ou rubricas.

Para saber quais seriam estes domínios, foi feita uma consulta ao dicionário. Apenas o substantivo-suporte *ordem* possui um uso especial na rubrica *biologia*, de modo que as relações que continham o hiperônimo *ordem*, sem complemento, não foram descartadas. Além de *ordem*, foram também consideradas as palavras *problema* – que pode ser considerada um substantivo-pleno na área de saúde – e *matéria* – que apareceu algumas vezes como sinônimo de disciplina, também sendo considerada um substantivo pleno. É importante salientar, porém, que, no caso de relações extraídas de um corpus não-específico quanto ao domínio, esta solução não é possível, sendo necessário então algum outro método para a determinação dos substantivos-suporte[24].

A lista de Marques (1995) de "substantivos de sentido geral" é composta por uma série de substantivos considerados altamente polissêmicos. A lista é baseada em uma parte do corpus do projeto NURC (Projeto de Estudo Conjunto e Coordenado da Norma Urbana Oral e Culta), provenientes de entrevistas realizadas na cidade do Rio de Janeiro. As entrevistas tratavam de temas específicos, como *política*, *ensino*, *vestuário*, etc, e foram considerados substantivos gerais aqueles de sentido geral que não têm vínculos com temas

específicos do NURC. Alguns substantivos da lista de Marques foram selecionados manualmente para serem filtrados, e também foram acrescentados os substantivos *itens*, *expressões*, *tema*, *informações* e *noções*. A lista final de substantivos que, quando apareceram exercendo a função de hiperônimos, acarretaram em exclusão de relações, engloba, portanto, (i) os substantivos-suporte descritos em Oliveira (2006); (ii) um subconjunto dos substantivos gerais descritos em Marques (1995) e (iii) alguns outros considerados gerais derivados de observação no corpus[25] (quadro 4).

âmbito, área, aspecto, assunto, base, campo, caráter, coisa, componente, cunho, dificuldade, dimensão, efeito, elemento, esfera, fator, forma, idéia, lado, maneira, modo, natureza, necessidade, nível, palavra, panorama, papel, parte, perspectiva, plano, ponto, quadro, questão, sentido, situação, termo, tipo, tom, itens, expressões, tema , informações, noções

Quadro 4: Substantivos gerais eliminados

6.2.2. Filtros 2 e 3: adjetivos e pronomes

A fim de diminuir os erros da categoria 2, relativos principalmente à "dependência contextual" de algumas relações, foram aplicados dois filtros: um para eliminação de pronomes dêiticos e outro para eliminação de alguns adjetivos.

6.2.2.1. Filtro de adjetivos

Hearst (1998) comenta que eliminou, nos seus resultados, adjetivos "comparativos", como *importante* e *menor*. Porém, embora a noção de adjetivo comparativo seja intuitivamente clara, não temos conhecimento, para a língua portuguesa, de uma lista de tais adjetivos que seja facilmente aplicada. Observando as relações extraídas, notamos que os adjetivos pré-nominais muito freqüentemente poderiam ser eliminados sem prejuízo significativo da informação, contribuindo para um caráter mais generalizador – menos contextual - do sintagma hiperônimo.

[24] É possível, por exemplo, utilizar o modelo de espaço vetorial empregado em Oliveira (2006).

capivara< ***grande*** *mamífero* → *capivara* < *mamífer*o

De uma perspectiva lingüística, a observação é compatível com a distinção entre adjetivos denotativos e predicativos: os primeiros acrescentam propriedades semânticas às propriedades da expressão nominal a que se referem; os últimos atribuem propriedades semânticas ao referente da expressão nominal modificada, acarretando em uma leitura proposicional (Lobato, 1993). Do ponto de vista formal, os adjetivos denotativos raramente aparecem em posição pré-nominal (Basílio et al., 2003), o que significa que, eliminando os adjetivos pré-nominais, corremos um risco muito pequeno de eliminar adjetivos que contribuem para a especificação do referente. Porém, se essas observações se aplicam perfeitamente no caso dos sintagmas hiperônimos, o mesmo não pode ser dito quanto aos sintagmas hipônimos. A diferença se deve à ambigüidade de determinadas relações hiper-hipo, que ora se referem apenas ao núcleo do sintagma hipônimo (e então a eliminação do adjetivo pré-nominal é bem-vinda), como em (e), ora se referem ao sintagma completo, incluídas as especificações decorrentes do adjetivo, como em (f) e (g), e ora são ambíguas (h).

(e) **pequenos** roubos<delinqüência

(f) **baixo** rendimento escolar<alterações comportamentais

(g) **menor** uso de intervenções obstétricas<efeitos benéficos de o suporte emocional no parto

(h) ***maior*** *consumo de leite*<*hábitos alimentares* .

Deste modo, no caso dos sintagmas hipônimos, por não ter, no momento, como identificar o referente exato do hiperônimo, optei por eliminar as relações iniciadas com adjetivos, com um pequeno sacrifício da abrangência em detrimento da precisão.

Porém, a apenas eliminação de adjetivos pré-nominais não é suficiente para levar a uma maior precisão nos resultados, pois ainda permanecem relações como

(i) arroz<alimentos **básicos**

[25] Embora os critérios para a escolha dos substantivos gerais tenham sido muito pouco

em que o adjetivo pós-nominal pode ser eliminado em nome de uma maior generalização. Foram excluídos então os "adjetivos gerais" de alta freqüência no corpus: a partir de uma lista com os 100 adjetivos mais freqüentes no corpus, separei, manualmente, aqueles de caráter geral – como *leve*, *grande*, *importante* –, dos de caráter específico do corpus – *humano*, *social*, *materno* – e (i) eliminei os adjetivos gerais e tudo o que estava à sua direita, na categoria Hiper; e (ii) eliminei toda a relação extraída em que o adjetivo está no sintagma hipônimo. O quadro 5 contém os adjetivos freqüentes que foram eliminados, por serem adjetivos "gerais".

amplo, anterior, básico, capaz, central, comum, diferente, difícil, direto, disponível, diverso, especial, específico, externo, freqüente, fundamental, geral, grande, gravíssimo, importante, inferior, inicial, maior, melhor, menor, múltiplo, necessário, normal, novo, pequeno, positivo, possível, presente, primeiro, próprio, relativo, responsável, seguinte, segundo, semelhante, significativo, simples, superior, total, último

Quadro 5: Adjetivos mais freqüentes e de caráter geral

O quadro 6 exemplifica o processo de filtro dos adjetivos, e a lista com os 100 adjetivos mais freqüentes está no anexo 1.

Relação original	**Tipo de filtro**	**relação final**
baixo rendimento escolar<alterações comportamentais	ADJ pré-nominal no Hipo	Eliminada
imperador José I da Áustria<personagens *importantes* de a história ocidental	ADJ freqüente /genérico no Hiper	imperador José I da Áustria<personagens
colesterol *alto*<problemas	ADJ freqüente /genérico no Hipo	Eliminada

Quadro 6: Exemplos da aplicação do filtro de adjetivos

Embora, ao menos no corpus utilizado, haja alguma sobreposição entre os adjetivos eliminados no filtro pré-nominal e os eliminados no filtro adjetivo genérico, preferi manter distinção entre as duas etapas, já que, por exemplo, em (j), *inexorável* deve ser eliminado (e de fato é, com o filtro de adjetivo pré-nominal), mas dificilmente apareceria em uma lista de adjetivos freqüentes.

"automáticos", eles se mostraram funcionais.

(j) perda de memória<**inexorável** deterioração de as funções cerebrais

Por fim, vale ressaltar que na busca por uma maior generalização dos termos muitas vezes especificações importantes se perdem, como mostra o quadro (7):

Relação original	Relação pós-filtro
leite desnatado<laticínios de **baixo** teor de gordura	leite desnatado<laticínios
favelas<áreas de **difícil acesso**	favelas<áreas
alcoólatras<pessoas com **baixa** imunidade	alcoólatras<pessoas
náusea<eventos freqüentes **em a gravidez**	náusea<eventos

Quadro 7: Exemplos de relações que perderam especificidades com o filtro ADJ

6.2.2.2. Filtro de pronomes dêiticos

O segundo filtro aplicado tem como objetivo eliminar pronomes dêiticos, como "meu", "seu", etc. As relações que contêm pronomes dêiticos não são excluídas - são alteradas para que a relação se mantenha, mas sem a referência ao contexto, como ilustram (k) e (l) :

(k) broncodiltadores < medicamentos prescritos **por seu** médico

(l) broncodilatadores < medicamentos prescritos **por** médico

6.3. Novos resultados

Após a aplicação dos filtros, o número de relações extraídas caiu de 2244 para 1937, isto é, pouco menos de 2% das relações foi eliminada. Das 1937 relações, 430 foram avaliadas manualmente. Os novos resultados estão na tabela 9.

Classificação	relações COM filtro	relações SEM filtro
3	**349 (81%)**	320 (73.4%)
2	**28 (6.5%)**	15 (3.4%)
1-	**20 (4.6%)**	70 (16%)
0	**33 (7.6%)**	31 (7.1%)

Tabela 9: Resultados da validação após aplicação dos filtros

A comparação dos resultados antes e depois da aplicação de filtros indica que a eliminação dos substantivos e adjetivos genéricos aumentou em 7% a

precisão dos resultados da categoria 3 (corretos), que agora correspondem a 81% das relações extraídas – e um grande declínio das relações classificadas como 1 – de 16% para 4.6%. Houve também uma pequena melhora nas relações classificadas como 2.

Com relação às relações erradas, classificadas como 0, cabe observar que, muitas vezes, o "erro" está no texto do corpus, e não é decorrente de problemas na metodologia empregada. Na frase abaixo, por exemplo,

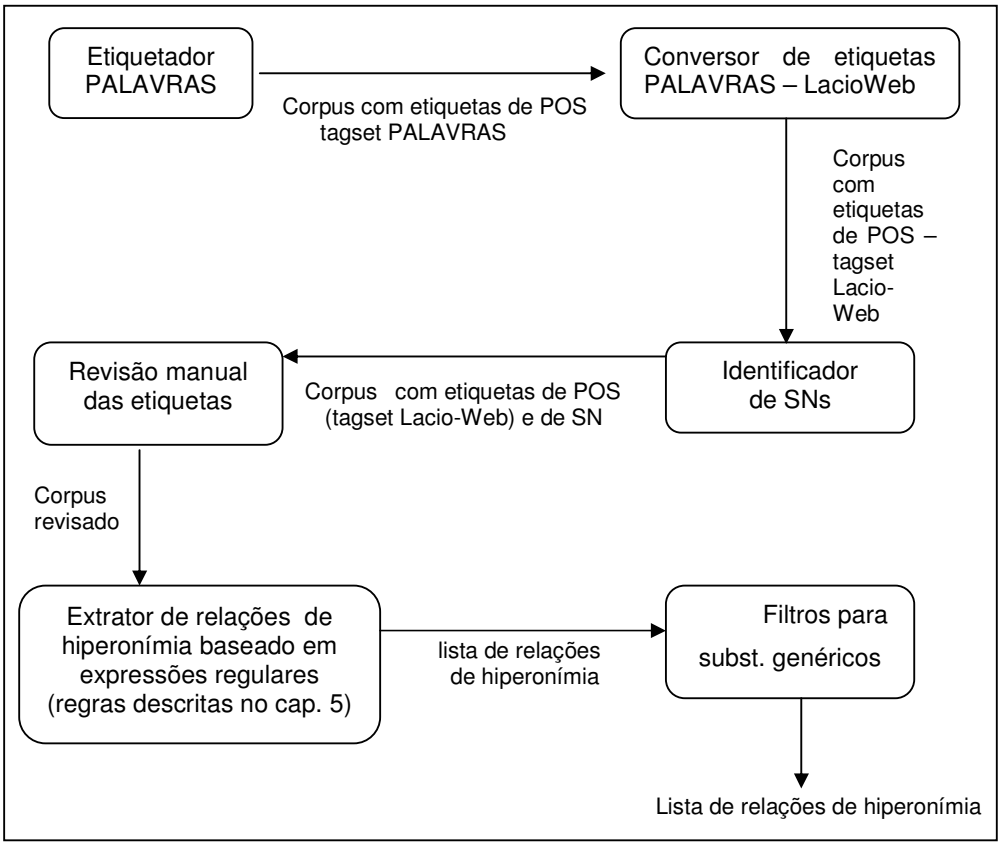
Entre os idosos examinados, houve alguns participantes que, por problemas clínicos, ***tais como*** *visão, audição, entre outros, não conseguiram completar...*

são extraídas as relações (m) e (n), o que aponta para algumas limitações quanto ao uso de corpus neste tipo de tarefa.

(m) visão<problemas clínicos

(n) audição<problemas clínicos

Com a incorporação dos filtros, o processo de extração de relações de hiperonímia no corpus está descrito no quadro 8.



Quadro 8: Processo de extração de relações de hiperonímia

6.4.
Generalização e comparação dos resultados

Com o objetivo de verificar se a metodologia empregada – especialmente os filtros – possui algum poder generalizador, obtendo sucesso não apenas no corpus específico em que foi aplicada, mas em qualquer corpus, todo o processo de identificação e extração de relações foi refeito em um pequeno corpus "genérico": uma amostra de 4862 sentenças (142.258 palavras) do corpus CETENFolha (Aires e Aluísio, 2001), um corpus que contém textos do jornal *Folha de São Paulo* do ano de 1994 e textos em registro didático, epistolar e redações de alunos. O corpus passou por 4 etapas de processamento, descritas no quadro 8.

Uma amostra de 527 relações foi analisada manualmente, e os resultados estão na tabela 10.

Classificação	Qtd de relações
3	397 (75%)
2	20 (3.7%)
1	32 (6%)
0	78 (14.8%)

Tabela 10: Resultados com o corpus genérico

Embora o índice de acertos (75%) seja inferior aos resultados obtidos com o corpus de saúde (81%), é importante lembrar que, neste momento, não houve uma eliminação prévia de erros "sintáticos", isto é, de erros decorrentes de ambigüidade na identificação de relações que contêm sintagmas preposicionais ou orações encaixadas. A metodologia foi utilizada nos resultados "brutos" das extrações. Daí, provavelmente, o grande aumento das relações classificadas como "erro" (categoria 0): de 7% (corpus saúde) para 14% (corpus genérico). Ainda assim, os resultados continuam superiores aos de Hearst (1998) e Cederberg e Widdows (2003), embora inferiores aos de Morin e Jacquemin (2004), como indica a tabela 11. Porém, como já comentado no final da seção 4.2.2, a comparação – principalmente com o trabalho de Morin e Jacquemin, deve ser vista com ressalvas, visto a forma de avaliação e o tipo de corpus serem diferentes.

	Amostra CETEN-Folha	Hearst[26] (1998)	Cederberg e Widdows (2003)	Morin e Jacquemin (2004)[27]
acertos	**397 (75%)**	104 (63%)	166 (64%)	286 (81%)
Total de relações avaliadas	**527**	166	260	353

Tabela 11: Comparação dos resultados

Uma observação interessante é a grande quantidade de relações que contêm nomes próprios (cerca de 52% de todo o corpus) e que receberam classificação 3 (cerca de 43%). Como o corpus que serviu de base para este último experimento é um corpus composto majoritariamente por textos jornalísticos, já era de se esperar um volume maior de nomes próprios, tanto de pessoas como de empresas e lugares. Uma possível explicação para o grande número de acertos envolvendo essa categoria está na própria estrutura dos nomes próprios: como são uma entidade única – um único *token* – não estão sujeitos aos erros de segmentação decorrentes da ambigüidade do SPrep. Por outro lado, a possibilidade de acerto é totalmente dependente de uma segmentação correta dos nomes próprios, tarefa que ainda apresenta desafios para a área de PLN (Mani e MacMillan, 1996; McDonald, 1996).

A comparação com os resultados obtidos em outros trabalhos demonstra que a metodologia empregada, lingüisticamente motivada, embora simples, foi bastante eficaz. Porém, é importante lembrar que o alto grau de subjetividade da tarefa de avaliação compromete o rigor da comparação.

Percebi, por exemplo, que alguns substantivos, embora não se encaixassem nas classes de genéricos e/ou suporte, também deveriam ser eliminados, por seu caráter transitivo[28]:

[26] Os resultados de Hearst (1998) referem-se apenas às relações extraídas com o padrão "e outros".

[27] Os resultados de Morin e Jacquemin (2004) referem-se apenas às relações extraídas com os padrões "tel que", "comme", "tel" e "et/ou de autre"

[28] Tais substantivos coincidem parcialmente com a descrição de substantivos *relacionais* feita por Bechara (1999): substantivos que não fazem referência a indivíduos, mas expressam relações entre indivíduos. Substantivos relacionais englobariam termos de parentesco como *pai*, *tio*, *irmão* (e *amigo*, *colega*, etc); e outros como *pátria* (em oposição a *país*), pois *pátria* está sempre relacionado a alguém, do mesmo modo que *mascote* (em oposição a *cão*), pois o *mascote* pressupõe um dono – diferentemente de *cão*. Bechara inclui ainda no grupo dos substantivos relacionais "nomes de partes do corpo e aqueles que aludem a partes constitutivas de uma

X < concorrente;
X < adversário
X < marido / pai/ esposa/ irmão.
X < parceiro

Tais relações foram consideradas categoria 1, isto é, relações muito gerais para serem úteis. Hearst (1998), porém, considera - erradamente, acredito – a relação

Nippon < partner

uma relação útil. E assim voltamos à fragilidade da forma de validação empregada, com o julgamento humano. Outras relações que apareceram no corpus também são de julgamento difícil, como

avião < peça feita com dobradura
alça de sutiã < lingerie ,

que foram classificadas como 1 e 0, respectivamente, evidenciando a opção por uma validação "conservadora".

Por fim, destacamos ainda que a quantidade de relações analisadas aqui foi superior a dos demais trabalhos (excetuando-se Morin e Jacquemin, 2004), o que também contribui para o caráter desigual da comparação. O quadro 9 apresenta um resumo comparativo entre este trabalho e os de Hearst (1998), Cederberg e Widdows (2003) e Morin e Jacquemin (2004).

É bastante curioso métodos simples como os empregados neste trabalho e em Morin e Jacquemin (2004) obtenham resultados melhores que o de Cederberg e Widdows, que testam uma combinação sofisticada de padrões baseados em expressões regulares e cálculos estatísticos. Credito o bom desempenho das regras que utilizei aos pequenos ajustes lingüísticos relacionados, principalmente, ao sintagma preposicionado, com a utilização das estruturas HHiper e HHipo. Além disso, acrescentei dois outros padrões ("tipos de" e "chamado") que apresentaram um alto grau de precisão.

---

entidade, física ou abstratamente considerada", como *braços da mulher*, *face do problema*, *galho da árvore* (Bechara, 1999:455).

	Corpus	**Qtde de relações analisadas**	**% de acertos**	**Técnica utilizada**
Hearst (1998)	6 meses de jornal *The New York Times*	166 relações (padrão "*e outros*")	63% (padrão "*e outros*"	Regras baseadas em expressões regulares
Cederberg e Widdows (2003)	430.000 palavras British National Corpus – corpus diversificado	260	64%	Regras baseadas em expressões regulares e cálculos estatísticos
Morin e Jacquemin (2004)	427.482 palavras domínio alimentos/agricultura resumos de artigos científicos (média de 316 palavras por resumo)	17 (padrão "*e outros*"	59% (padrão "*e outros*"	Regras baseadas em expressões regulares descobertas automaticamente
		353 (padrões "e outros", "como/tais como"	81% (padrões "e outros", "como/tais como"	
		1216 (todos os padrões)	**82%**	
Freitas (2007)	1.846.502 palavras corpus diversificado, majoritariamente jornalístico	527	75%	Regras baseadas em expressões regulares

Quadro 9: Resumo comparativo

Já os resultados de Morin e Jacquemin (2004) são de difícil interpretação, principalmente devido ao corpus utilizado. Como se trata de um corpus de um domínio restrito, que contém apenas resumos de textos técnicos, é possível que o material lingüístico seja mais simples em termos de estruturas sintáticas, com uma menor ocorrência de sintagmas preposicionados, por exemplo, o que pode levar a uma baixa freqüência de estruturas ambíguas – problema já notado em Hearst que nem chega a ser comentado pelos autores.

Por fim, lembro que a subjetividade da tarefa de avaliação também interfere na exatidão da comparação, bem como as diferentes condições em que os trabalhos foram feitos, de modo que a comparação deve ser vista com cautela.

7
Produzindo conhecimento novo: a realização de inferências

A maioria dos trabalhos que envolve a extração de relações de hiponímia não utiliza os resultados dessa extração para a realização de inferências. Uma possível explicação para esse descarte é a grande quantidade de erros produzidos, principalmente quando se trata de relações extraídas de corpus gerais quanto ao domínio, como é o caso de corpus de textos jornalísticos. Kilgarriff (2003) se opõe à utilização de tesauros baseados em palavras (com relações extraídas diretamente do corpus) como ontologias na IA justamente por ser a realização de inferências – raciocínio fundamental em ontologias e em IA – um processo baseado em conceito, em significado. Defensor de uma perspectiva relativista com relação ao significado, Kilgarriff é consciente das imprecisões dos significados das palavras, e por isso argumenta que inferências são um problema para trabalhos baseados em corpus. Um exemplo: em uma ontologia baseada em corpus – e em palavras – , teríamos que *tucanos* são *aves*. Poderíamos encontrar, também, que alguns *políticos* são *tucanos*, mas não gostaríamos de inferir que alguns *políticos* são *aves*[29]. De fato, este é um passo delicado, uma vez que inferências pressupõem um significado fixo e estável das palavras. Porém, em favor de uma ontologia baseada em palavras, argumento que o fato de nos apoiarmos em um corpus específico de domínio deve evitar a ocorrência de situações como a descrita por Kilgarriff. Para tanto, invoco a restrição "*one sense per discourse*" (Yarowsky 1995), segundo a qual o significado de uma dada palavra é altamente consistente em um determinado texto. Como o corpus de trabalho é específico de domínio, espero que a restrição possa ser ampliada de "texto" para "domínio".

Em um primeiro cruzamento das informações, isto é, o agrupamento das relações extraídas com as regras de identificação de hiperonímia, foi observado

29 O exemplo original é "However it cannot be the word cat that maps directly to the ontology, as some cats are jazz musicians, and we do not wish to infer that they are furry." (2003:5)

um número excessivo de taxonomias[30] independentes que deveriam estar relacionadas – não havia conexão, por exemplo, entre as taxonomias de *sintomas*, *sintomas agudos* e *sintomas de gripe*, o que parece contra-intuitivo. A fim de relacionar as taxonomias, foi criada uma regra simples que gera, para sintagmas hiperônimos compostos por mais de um substantivo, um novo hiperônimo formado pelo substantivo núcleo do sintagma, chamada regra HiperN. Com isso, foram produzidas, automaticamente, as seguintes relações

sintomas agudos < *sintomas*

sintomas de gripe < *sintomas*

que então podem ser integradas à taxonomia de *sintomas* (figura 6, pg 99). Contudo, a aplicação da regra HiperN gera categorias hiperônimas indesejáveis nos seguintes casos:

(i) o hiperN criado é um substantivo deverbal, que carrega a transitividade do verbo e cuja utilização como hiperônimo causa estranheza justamente pela ausência do complemento. A figura 3, da taxonomia de *adoção,* ilustra esse caso;

(ii) os substantivos suporte/genéricos, eliminados pelos filtros, voltam a aparecer como hiperônimos. Por exemplo, para o sintagma *áreas de apoio* é criado o hiperônimo *áreas* .

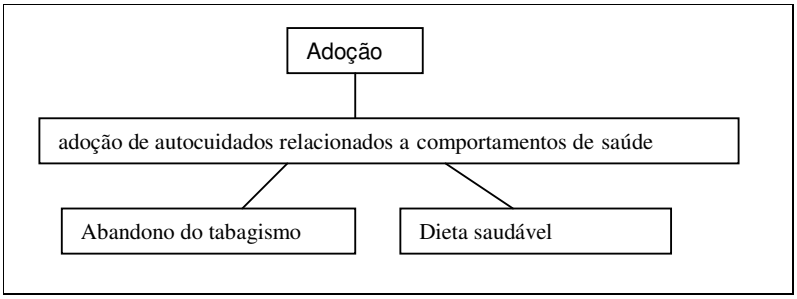


Figura 3: Taxonomia de adoção produzida pela regra "hiperN"

Com relação a (i), parece difícil eliminar o problema sem consultar informação morfossintática a respeito do nome. Já (ii), se, por um lado, é de resolução bem mais simples – basta reaplicar o filtro para eliminação dos substantivos suporte/genéricos – , por outro, envolve uma decisão teórica não tão

[30] Chamo de taxonomias o resultado do cruzamento das relações obtidas com a aplicação das regras.

simples: será mesmo que, aparecendo apenas como um substantivo hiperônimo "aglutinador", sob o qual se agrupam as diversas possibilidades de ocorrências e de significação, é desejável a eliminação do substantivo genérico? Em outras palavras: se mudarmos ligeiramente o foco de utilização da taxonomia – de caracterização de um domínio para o levantamento lexicográfico de um domínio, é desejável sua eliminação? Enfim, seria (ii) realmente um problema? Os substantivos genéricos, quando voltam a funcionar como hiperônimos, explicitam seus diversos contextos de uso, o que nos levou a não considerar esses casos como erros. A figura 4 ilustra a taxonomia de "área".

ÁREAS
—áreas de apoio
— —psicologia
— —saúde pública
— —terapia ocupacional
—áreas de conhecimento
— —astrofísica
— —cosmologia
— —física de partículas
—áreas do sistema nervoso central associadas ao medo
— —substância cinzenta periaquedutal dorsal
—áreas de repouso
— —camas
—áreas hiperendêmicas de doença meningocócica
— —cinturão da meningite
—áreas prioritárias
— —alimentação
— —educação
— —moradia
— —renda
— —saneamento
— —segurança
— — —fornecimento de proteção individual
— — —ventilação forçada
—áreas silvestres
— —florestas
— —regiões de cerrado

Figura 4: Taxonomia de *áreas*

Com o cruzamento das informações obtidas na extração dos padrões léxico-sintáticos, foram encontradas 420 taxonomias no domínio saúde. Dessas, cerca de 1/3 foi selecionada para avaliação manual. Uma primeira análise revelou um grande número de taxonomias com apenas dois níveis. Como o objetivo desta parte da avaliação é a análise da produção de inferências, a avaliação foi limitada apenas às taxonomias que possuem mais de dois níveis, isto é, taxonomias cujos resultados são diferentes dos resultados da aplicação das regras. Além disso,

dentre as taxonomias de dois ou mais níveis, havia taxonomias "artificiais", isto é, taxonomias cujo terceiro nível resultava da aplicação da regra HiperN. Uma vez que o objetivo dessa regra não é produzir inferências, mas sim agrupar taxonomias relacionadas (por exemplo, agrupar em uma única taxonomia *bois* e *cavalos*, que são hipônimos de *animais de grande porte*; e *gatos* e *cachorros*, que são hipônimos de *animais*, em uma taxonomia única, *animais*), também foram descartadas da avaliação as taxonomias com 3 ou mais níveis resultantes da aplicação da regra como ilustra a figura 5.

ALÉRGENOS
—alérgenos inalantes
— —ácaros
— —poeira doméstica

Figura 5: Taxonomia com inferência "artificial"

Com isso, das 188 taxonomias, sobraram 96 taxonomias para serem avaliadas manualmente.

Surpreendentemente, encontramos erros em apenas 9 taxonomias, num total de 90% de acertos, o que contradiz a posição de Kilgarriff de que não é possível a realização de inferências em trabalhos baseados em corpus. Por outro lado, esse alto índice de acertos se deve, em grande parte, à utilização de um domínio restrito e técnico, o que dá pouca margem à ocorrência de variações entre os significados. De fato, como já assinala Cruse (1986), o vocabulário científico é mais preciso que o vocabulário cotidiano. A figura 6 apresenta a taxonomia de *sintomas*.

Uma análise cuidadosa das taxonomias corretas revelou dados interessantes: algumas taxonomias ficaram muito grandes, principalmente aquelas cujo termo hiperônimo possuía, como um dos hipônimos, o termo *doenças* – o que está de acordo com o que se espera da representação de conhecimento da área de saúde. As taxonomias de *infecções*, *agravo*s e *complicações* ilustram este fato (anexos 2-4).

SINTOMAS
—agitação
—alterações em os batimentos cardíacos
—alterações visuais
—anorexia
—ânsias
—comprometimento de os rins
—coriza
—diarréia intermitente
—dificuldade
—dor de cabeça
—dor muscular
—dor
—dores de cabeça
—dores de estômago
—dores de garganta e de cabeça
—dores em o peito
—espirros
—estresse
—fadiga
—febre
—fígado
—hemorragias
— —epistaxe
— —gengivorragia
—icterícia
—infecção branda de o trato respiratório
—insatisfação com o trabalho
—linfadenopatia generalizada
—perda de peso
—problemas cardíacos
— —embolias
— —tromboses
—sintomas agudos
— —febre
—sintomas de gripe
— —conjuntivite
— —dor em o corpo
— —febre
—sintomas essencialmente agudos
— —cloracne
—sudorese noturna
—tontura
—tosses eventuais

Figura 6: Taxonomia de *sintomas*

Dos 9 erros encontrados, 6 são conseqüência de polissemia[31]. O quadro 10 ilustra os 6 casos, com a palavra indutora de erro em negrito.

[31] O termo polissemia é utilizado conforme descrito em Martins (1999): uma multiplicidade de usos que os falantes podem regularmente atribuir às palavras, manifestando sua capacidade de participar dos jogos de linguagem em que a palavra comparece.

MATERIALIDADES —água — —água sanitária —alimentos — —**açúcar** — — —*dextrana (?)	GULOSEIMAS —**açúcar** — —dextrana (?) —balas —café —enlatados	DETALHES —**efeitos colaterais** — —dor de cabeça — —erupções de a pele — —náusea — —*paralisia definitiva (?) — —vertigens
ASSOCIAÇÕES —associações científicas — —Sociedade Brasileira de Medicina Tropical —obesidade (?)	HÁBITOS —**drogas** — —antiinflamatórios(?) — —anti-retrovirais(?) — —bloqueadores de secreção ácida(?) — —cloroquina(?)	FENÔMENOS —**drogas** — —antiinflamatórios(?) — —anti-retrovirais(?) — —bloqueadores de secreção ácida(?) — —cloroquina(?)

Quadro 10: Taxonomias que produziram erros em decorrência de poslissemia

Nos exemplos das taxonomias de *hábitos* e *fenômenos* o problema da inferência está em *droga*, que pode ser compreendida como um *fenômeno social*, como *hábito* ou como *substância*. A figura 7 mostra a interseção entre os três usos de *droga*. O que o sistema faz é "exportar" os hipônimos de *droga_substância*, que não possuem hiperônimo no corpus, para os hiperônimos *hábitos* e *fenômenos*.

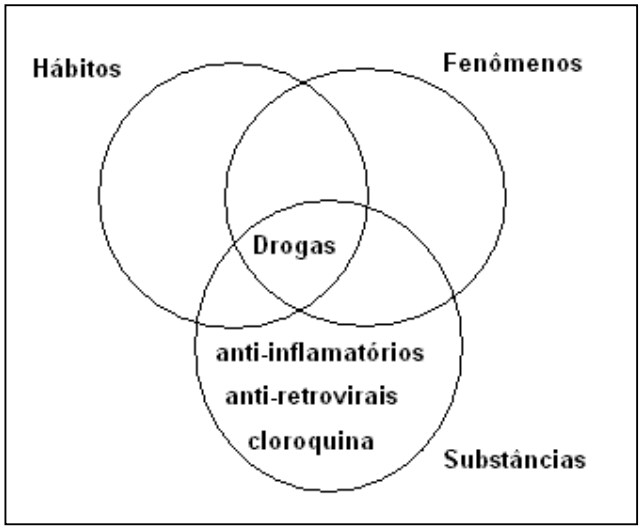


Figura 7: Diferentes contextos de uso de *drogas*

Já nos exemplos de *guloseimas* e *materialidades* há uma clara evidência de diferença quanto aos registros utilizados – do ponto de vista técnico, *dextrana* é um tipo de *açúcar*; do ponto de vista da linguagem ordinária, *açúcar* é uma

guloseima e um *alimento*. Embora o corpus seja de um domínio técnico, ele também possui textos de divulgação, o que justifica este tipo de ocorrência. Aliás, é justamente a presença de textos não tão técnicos no corpus que possibilita grande parte dos acertos, como mostra o exemplo da figura 8. A relação entre *mosquitos flebótomos* e *artrópodes* dificilmente seria explicitada em algum texto, pois estão em níveis diferentes de especialidade. E, de fato, uma busca no *Google* pela expressão "*mosquitos flebótomos são artrópodes*" não retornou nenhum documento – o que também reforça a dificuldade de avaliação deste tipo de tarefa, como já discutido no capítulo 4.

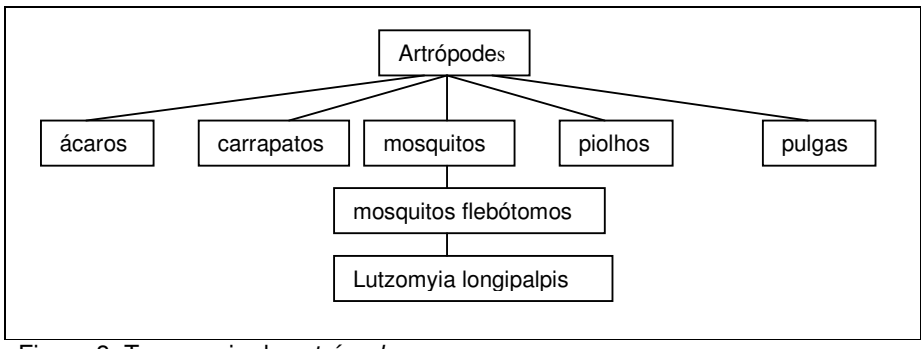


Figura 8: Taxonomia de *artrópodes*

Dos 3 outros erros encontrados na avaliação das taxonomias, um é de correção fácil: o hiperônimo é *palavra-chave*, que pode ser incluído no filtro para eliminação dos substantivos gerais. Os outros dois erros são decorrência da regra HiperN: em um caso, o hiperônimo é o termo *conjunto* funcionando como um quantificador ("conjunto de"), que talvez também possa ser incorporado em um filtro (figura 9); no outro erro, o problema está no fato do corpus não possuir etiquetas consistentes para expressões multi-vocabulares (EMVs) nominais. Deste modo, para a EMV *estilo de vida* é criado o hiperônimo "estilo" (figura 10).

CONJUNTO
—conjunto de características prévias à menopausa
— —comportamentos desfavoráveis à saúde
—conjunto de valores éticos
— —cidadania
— —eqüidade
— —solidariedade
— —vida

Figura 9: Taxonomia de *conjunto*

ESTILOS
—estilo de vida
— —hábito
—estilos de vida não saudáveis
— —abuso de substâncias
— —hábitos alimentares nocivos
— —sedentarismo

Figura 10: Taxonomia de *estilos*

Algumas vezes as taxonomias deixaram de exibir relações esperadas entre os termos. Na taxonomia de *infecções*, por exemplo (figura 11), *diarréia* e *bronquite* estão diretamente ligadas ao nó mais alto *infecções*, ocupando o mesmo nível de *infecções agudas*, *infecções bacterianas*, *infecções cutâneas* e *infecções virais*. Porém, para que o paralelismo entre os nós fosse mantido, o mais correto seria que *diarréia* e *bronquite* estivessem subordinadas a categorias como *infecção intestinal* e *infecção respiratória*, mas tais categorias não "emergiram" do corpus.

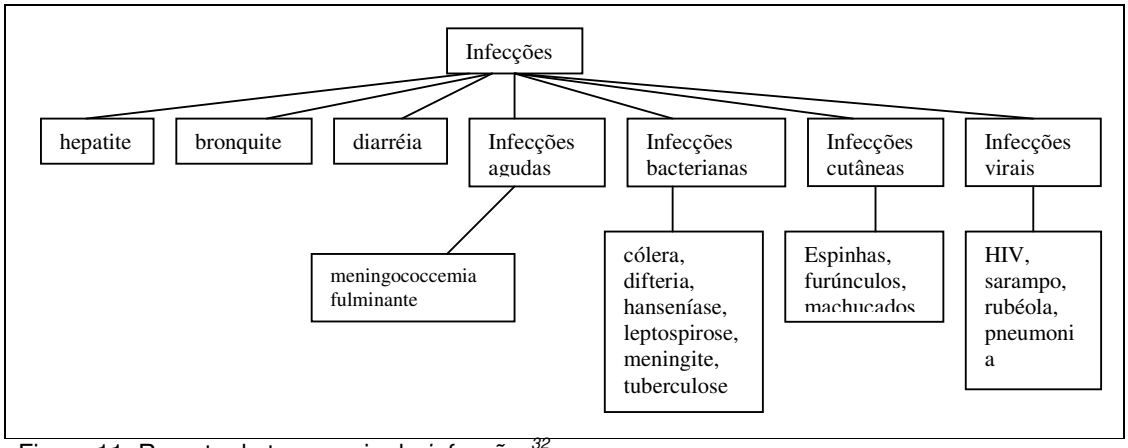


Figura 11: Recorte da taxonomia de *infecções*[32]

O mesmo pode ser observado com a taxonomia de *objetos* (figura 12): era de se esperar que *faca* aparecesse como subordinado ao hiperônimo *talheres*, o que não aconteceu. Esses casos, porém, não foram considerados erros, mas decorrência da característica das taxonomias naturais de freqüentemente não apresentarem nós em todos os níveis, já apontada por Cruse (1986), o que só

[32] A taxonomia completa de infecções está no anexo 2

reforça o caráter híbrido das taxonomias construídas. Por outro lado, as lacunas lexicais a que Cruse se refere seriam conseqüência de conceitos hiperônimos não lexicalizados na língua. No caso de *infecção*, por exemplo, o problema é de outra natureza: o hiperônimo em questão existe na língua, mas ou não foi capturado pelas regras de extração ou não existia no corpus. Porém, em favor da metodologia apresentada, argumento que mesmo na Wordnet (Fellbaum, 1998), construída manualmente, esta situação ocorre (Lin e Pantel, 2002).

Outra característica das taxonomias naturais observada aqui foi o número reduzido de níveis: a maioria das taxonomias não teve mais que 3 níveis, o que também está de acordo com o relatado na literatura (Cruse, 1986; Lyons, 1980).

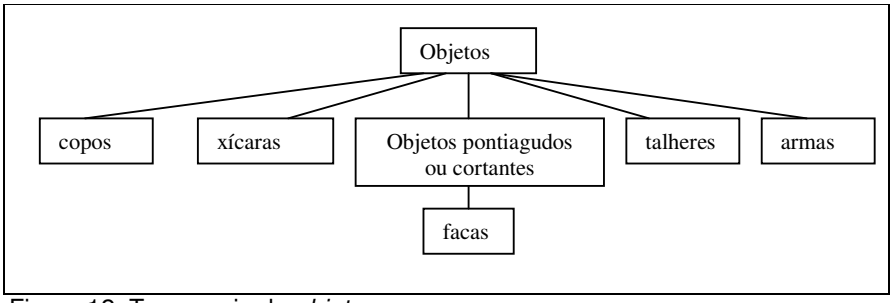


Figura 12: Taxonomia de *objetos*

Por fim, o cruzamento dos dados para a inferência acabou possibilitando a realização de heranças múltiplas, característica que diz respeito à localização de um termo em múltiplas posições na taxonomia, evidenciando sua multiplicidade de significados. A possibilidade de herança múltipla tem conseqüências no formato geral da ontologia pois, ao invés de estruturas de árvore, o conhecimento representado passa a ter a forma de um grafo acíclico, no qual alguns nós possuem mais de um pai. O termo *fumo* , por exemplo, é ao mesmo tempo *droga estimulante* e *fator de risco*; *frituras* são simultaneamente *alimentos gordurosos* e *guloseimas.*

Embora a estrutura de grafo seja a representação mais fiel das relações entre as palavras, às vezes esta representação pode ficar muito complexa. Por outro lado, a herança múltipla pode ser mais facilmente observada se simplesmente invertemos a forma de visualização da taxonomia. Em outras palavras: os exemplos analisados até agora mostram a taxonomia em seu formato "original", isto é, uma taxonomia *top-down*. Existe, contudo, uma outra forma de observar as

relações produzidas que pode ser de grande utilidade para a lexicografia. Se os termos são gerados de maneira *bottom-up*, do mais específico para o mais geral, características bastante interessantes ficam realçadas. De certa maneira, os resultados, principalmente nas relações de apenas um nível, se assemelham aos apresentados nas wordnets, ainda que sem as definições. Porém, muitas vezes a própria relação de hiponímia, principalmente quando existe apenas um nível, pode funcionar como uma espécie de definição, como mostra o quadro 11.

ala desaminase < enzima
difteria < infecções bacterianas
Dinamarca < países europeus
dióxido de nitrogênio < gases poluentes
dispnéia < complicações respiratórias
doença falciforme < hemoglobinopatias
dor no corpo < sintomas de gripe
efisema < complicações respiratórias
implantação de pontes em artérias coronárias < procedimentos cirúrgicos
Instituto Butantan < instituições públicas
Institutos Manguinhos < estabelecimentos diretamente ligados à área de epidemiologia
meprobamato < droga
microbiologistas < cientistas de a área biológica
MSX 1 < gene
multimistura < suplemento alimentar
Mycobacterium tuberculosis < bactéria
privação de água ou alimento < maus-tratos
ipês-rosas < espécies nativas brasileiras
roturas himenais < lesões genitais
ruas < espaços urbanos públicos
rubéola < infecções virais
ruptura de o diafragma < complicações respiratórias
saturação da transferrina < indicadores bioquímicos de a situação orgânica de ferro
tranqüilizantes < drogas prescritas por médicos
transparência < recursos audiovisuais
trens < meios de transporte
urocultura < exames
uso de anticoncepcionais < fatores individuais de risco

Quadro 11: Resultados da taxonomia no formato *bottom-up* para relações de 1 nível

Além da aparência definitória nos casos de taxonomias com apenas um nível, outro aspecto interessante da visualização *bottom-up* é a explicitação dos diversos contextos de uso dos termos. O quadro 12 apresenta alguns resultados de taxonomias com mais de um hiperônimo[33]:

[33] No quadro, como há uma "inversão" na visualização, o termo em negrito é o hipônimo, e os que estão abaixo dele são os hiperômimos.

amendoim —componentes de um suplemento alimentar chamado multimistura —grãos **São Paulo** —cidade —estados —metrópoles —município de grande porte **tuberculose** —condições crônicas —doenças — —agravos à saúde — —desfechos — —doenças crônicas — —intercorrências —doenças de transmissão respiratória —infecções bacterianas —pneumopatias **saliva** —fluidos —secreções —secreções de as vias aéreas **arroz** —alimentos — —materialidades —culturas temporárias —gramíneas — —forrageiras **colesterol HDL** —colesterol — —nutrientes — —problemas	**álcool** —drogas estimulantes —drogas sedativas —substâncias tóxicas **sarampo** —complicações —doenças febris —doenças infecciosas —infecções —infecções raras em adultos —infecções virais —infecções virais sistêmicas **ansiedade** —distúrbios —fatores psicológicos —itens sobre a emoção —problemas considerados da esfera emocional **oligopepetidases** —enzimas — —substâncias **diarréias** —complicações —infecções —patologias típicas do subdesenvolvimento —distúrbios —doenças — —agravos à saúde — —desfechos — —doenças crônicas — —intercorrências —doenças tipicamente relacionadas com o lixo	**dor de cabeça** —distúrbios —efeitos colaterais — —detalhes —efeitos desagradáveis —sintomas **virilha** —dobras de pele —partes de o corpo **Brasil** —país endêmico —países —países americanos —países da América Latina —países em desenvolvimento **cólera** —doenças — —agravos à saúde — —desfechos — —doenças crônicas — —intercorrências —doenças infecciosas intestinais —infecções bacterianas **roubos** —condutas anti-sociais —delitos **sangue** —fluidos corporais potencialmente infectantes —materiais biológicos ricos em células

Quadro 12: Resultados de visualização *bottom-up* para taxonomias com mais de um hiperônimo

7.1. Inferências em um corpus genérico

A fim de verificar se o alto índice de acertos obtido na realização de inferências foi conseqüência da utilização de um corpus de domínio específico, o mesmo processo de cruzamento de dados foi realizado com a amostra do corpus CETENFolha, de cerca de 142.00 palavras. Foram produzidas 920 taxonomias.

Uma primeira observação diz respeito ao alto número de taxonomias, principalmente se considerarmos que o corpus de saúde, com quase 2 milhões de palavras, produziu 420 taxonomias. Essa proliferação excessiva de taxonomias no corpus geral é conseqüência de dois fatores: (ii) o caráter geral do corpus CETENFolha, que trata de uma vasta gama de assuntos; (i) a "ausência" de inferências, isto é, grande parte das taxonomias possui apenas 2 níveis, o que corresponde ao resultado das regras de extração de hiperonímia. Por outro lado, esses resultados não chegam a ser surpreendentes, visto a presença de poucos níveis de profundidade ser uma característica das taxonomias naturais, como já observaram Cruse (1986) e Lyons (1980).

Outro aspecto que diferencia a ontologia de domínio e a ontologia geral é a presença, na última, de taxonomias com muitos hipônimos, unificadas por termos que acabaram funcionando como termos genéricos em um contexto jornalístico, como *produtos* (184 hipônimos), *utensílios* (137 hipônimos), *profissionais* (104 hipônimos), *conceitos* (101 hipônimos), *instituições* (82 hipônimos); ou por termos cujos hipônimos são freqüentes e numerosos em jornal, como *países* (118 hipônimos) e *jogadores* (79 hipônimos). Nas maiores taxonomias – as de *produtos* e *utensílios* – , que são uma espécie de categoria "coringa", capazes de abrigar quase qualquer palavra, foram poucos os erros encontrados. No caso específico de *utensílios*, seu caráter abrangente se deve principalmente à presença de *objeto*, que também é bastante abrangente, como um dos hipônimos. A taxonomia de *conceitos* apresentou muitos erros, principalmente devido à natureza mais "abstrata" de *conceito*, que favorece a presença de polissemia. As demais taxonomias "gigantes" possuem poucos erros – e também poucos níveis – e são sobretudo categorias que abrigam nomes próprios, o que já é indicativo do potencial desta metodologia para a classificação semântica dessa classe de palavras (as taxonomias de *produtos*, *utensílios, países*, *profissionais*, *conceitos*, *instituições* e *jogadores* estão nos anexos 5-11).

Das 920 taxonomias produzidas, 234 foram avaliadas manualmente. Novamente, a análise foi limitada apenas às taxonomias que possuem mais de dois níveis. Com isso, sobraram 50 taxonomias para avaliação manual.

Os resultados mostram que, das 50 taxonomias, 20 (40%) possuem erros decorrentes da polissemia, em um quadro muito diferente dos resultados obtidos no corpus de saúde. Seguindo as previsões de Kilgarriff (2003), poucas

inferências produziram resultados satisfatórios. Não encontrei nenhum *Cat Stevens peludo*[34], mas me deparei com um *B.B. King* que é um *adorno fofo*, como mostra a figura 13. As figuras 14 e 15 exemplificam outros casos de polissemia (a palavra indutora de erro está em negrito).

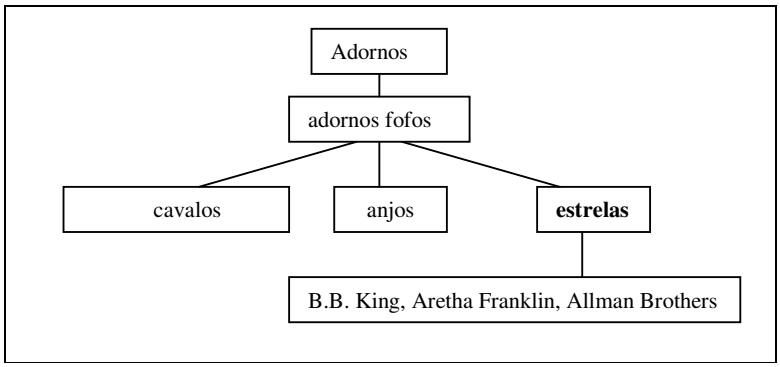


Figura 13: Taxonomia de *adornos*

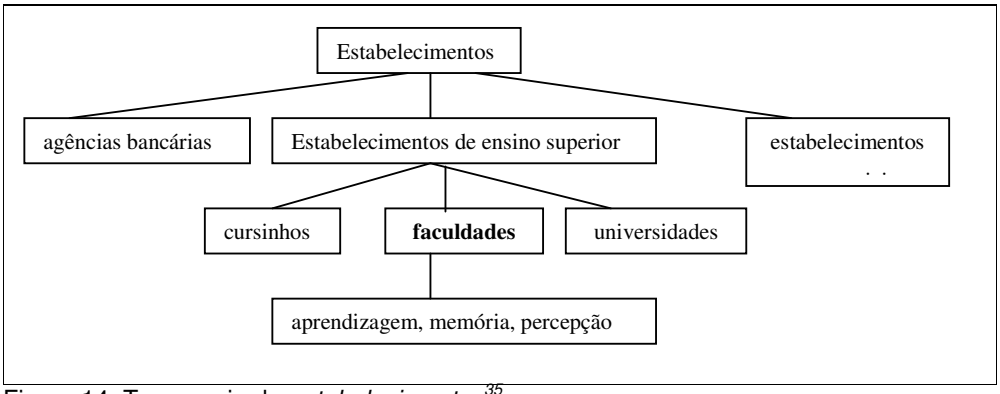


Figura 14: Taxonomia de *estabelecimentos*[35]

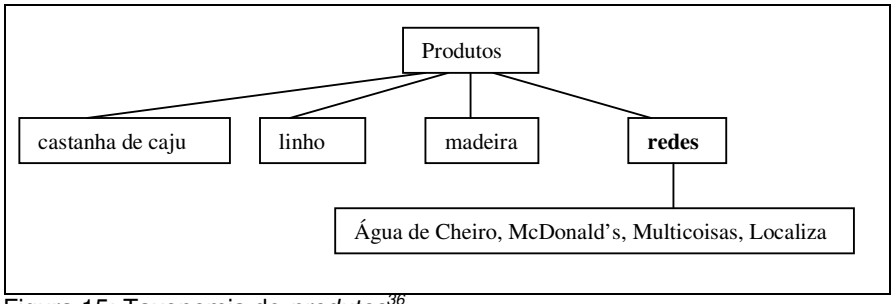


Figura 15: Taxonomia de *produtos*[36]

[34] Conferir a nota 5.

[35] A taxonomia completa de *estabelecimentos* está no anexo 10.

[36] A taxonomia completa de produtos está no anexo 11

De fato, em um corpus não específico, a polissemia é mais aparente, impedindo o caminho lógico das inferências. Fica patente, neste caso, a discrepância na aplicação de uma ferramenta lógica, precisa – as inferências – em um objeto assumidamente fluido – a língua cotidiana, com um vocabulário não específico. Some-se a isso o fato de que, no corpus de jornal, co-existem diferentes graus de formalidade e uma grande diversidade de assuntos, o que dificulta ainda mais as inferências, como é possível observar nos exemplos (a) (visualização top-down) e (b) (visualização bottom-up):

(a) *frutas*
—abacaxi
— —Banespa

(b) Asterix
—heróis
— —pilares da dramaturgia

Porém, se a produção de inferências não é possível em um corpus geral, a visualização dos resultados *em formato bottom-up* (sem as inferências, apenas com os resultados das regras) pode ser um auxílio para o lexicógrafo, justamente por evidenciar os diferentes contextos de uso das palavras. O quadro 13 ilustra algumas palavras e seus diferentes hiperônimos:

desenho	**carne**	**cólera**
—atividades	—alimentos	—doenças
—elementos visuais	—filés	—fatos psíquicos
—recursos plásticos	—produtos	—males obsoletos
—técnicas	—proteínas	—doenças causadas pela falta de condições sanitárias
milho		
—culturas	**hospitais**	**futebol**
—culturas anuais	—ambientes	—esporte
—espécies	—compradores institucionais	—jogo
—frutos	—entidades	—modalidades
—grãos	—locais públicos	—mundo infernal
—produtos	—serviços essenciais	—produto

Quadro 13: Visualização *top-down* de relações da amostra do CorpusCETENFolha

7.2. Nomes Próprios

Por fim, uma última observação com relação aos resultados diz respeito aos nomes próprios. Cerca de 10% do total de relações de hiponímia identificadas no corpus de saúde têm como elemento hipônimo um nome próprio.

Uma análise manual do material extraído revelou um alto grau de precisão – 98% de acertos em uma amostra de 100 relações. Tais resultados são encorajadores para a utilização das regras de identificação de hiponímia como auxiliares de sistemas de classificação semântica de nomes próprios. Uma das vantagens da utilização da técnica é justamente a possibilidade de lidar com a variação de sentido característica dessa classe de nomes. O exemplo de *Rio de Janeiro*, retirado do corpus, é uma boa ilustração:

Rio de Janeiro
-aglomerados urbanos
-capitais
-cidades
-estado

Nomes próprios costumam ser considerados, pela teoria lingüística, um fenômeno periférico, por não oferecerem contribuições relevantes sobre o funcionamento da estrutura da(s) língua(s). Talvez em conseqüência dessa desvalorização, imagina-se que sua identificação e classificação semântica automática seja uma tarefa simples, o que não corresponde à realidade. Por outro lado, o processamento dos nomes próprios é crucial na análise de textos, pois são unidades lingüísticas que aparecem com freqüência bastante significativa na língua.

Alguns trabalhos sobre identificação e classificação automática de nomes próprios fazem uso de listas de antropônimos e topônimos, ou de outras bases de conhecimento (Mani e MacMillan, 1996). Porém, tais listas costumam apresentar limitações, como a custosa elaboração manual, que acarreta em dificuldades de atualização e extensão e, freqüentemente, uma quantidade sempre insuficiente de nomes próprios. O fato de nomes próprios constituírem uma classe ainda mais "aberta" do que a dos substantivos comuns salienta a necessidade de atualização constante e, conseqüentemente, de metodologias capazes de acrescentar nomes – e suas classes semânticas – automaticamente.

O tratamento computacional de nomes próprios envolve duas tarefas: a segmentação dos nomes e, posteriormente, sua classificação semântica. Quanto à segmentação, o principal problema consiste em delimitar as fronteiras de um nome próprio.

(1) Philip B. Morris
(2) Juiz Nicolau dos Santos Neto
(3) Presidente da Câmara dos Vereadores Alcides Barroso

Em (1), a dificuldade consiste em impedir que o sistema reconhecedor interprete o ponto após a letra B como um ponto final, e conseqüentemente *Morris* como uma outra palavra, ao invés de integrante do único nome em questão. Em (2), o problema é o inverso: é preciso distinguir dois termos no sintagma: o substantivo comum *juiz* e o nome próprio *Nicolau dos Santos Neto*. Em (3), a dificuldade está na polissemia da construção: a segmentação pode feita em (i) *presidente* e (ii) *Câmara dos Vereadores Alcides Barroso*, ou em (i) *presidente*, (ii) *Câmara dos Vereadores* e (iii) *Alcides Barroso*, em que (i) e (iii) são co-referentes.

Como o corpus utilizado aqui já foi processado pelo etiquetador PALAVRAS (Bick, 2000), não foi preciso lidar a etapa de segmentação dos nomes próprios. Mas é importante lembrar que, no processo de revisão manual das etiquetas, houve também a preocupação de corrigir problemas decorrentes de erros de segmentação, o que certamente contribuiu para o grande número de acertos.

Já a classificação semântica de nomes próprios integra a área de Reconhecimento de Entidades Mencionadas (REM), cujo objetivo final é a identificação e classificação de palavras e expressões (chamadas entidades mencionadas) em determinadas categorias semânticas pré-definidas, como *pessoa, organização, localização, tempo, data, percentuais* e *expressões monetárias*, que, por sua vez, podem se subdividir: a categoria *localização*, por exemplo, pode englobar as subcategorias *localização geográfica* e *localização política* e/ou *administrativa*.

Com a metodologia empregada aqui não existem rótulos semânticos pré-estabelecidos, mas apenas aqueles revelados no corpus. Neste ponto, uma desvantagem da metodologia é a dificuldade de comparação com outros classificadores semânticos; por outro lado, a abordagem proposta oferece mais possibilidades para que a polissemia – expressa pelas múltiplas faces de um mesmo nome próprio – apareça, como no exemplo de *Rio de Janeiro*. Uma abordagem que utilize a informação obtida com as regras de extração de

hipônimos e a compatibilize com categorias semânticas pré-definidas parece ser um caminho produtivo na pesquisa sobre o reconhecimento de entidades nomeadas. No anexo 12 estão alguns resultados de relações que envolvem nomes próprios no corpus de saúde.

7.2.1. Classificação semântica de nomes próprios em um corpus genérico

Se a realização de inferências foi pouco promissora com a utilização do corpus genérico, o mesmo não acontece com a classificação de nomes próprios. Como algumas "taxonomias gigantes" já indicavam, a grande quantidade de relações cujo hipônimo é um nome próprio é um indício de que a aplicação das regras pode ser uma estratégia eficaz para a o tratamento desta classe de nomes.

No corpus genérico, do total de 5267 relações de hiperonímia extraídas com as regras, 2418 (46%) – quase *metade* das relações – têm como hipônimo um nome próprio. É um número altíssimo, principalmente em comparação com os resultados do corpus de saúde, como mostra a tabela 12.

	Tamanho (em palavras)	Qtde de relações[37]	**Qtde de relações cujo hipônimo é um NPprop**
Corpus de Saúde	1.846.502	2.932	**10%**
Amostra do corpus CETENFolha	**142.258**	**5.217**	**46%**

Tabela 12: Comparação entre os corpora com relação aos nomes próprios

Das 2.418 relações com nomes próprios, aproximadamente 1/3 foi selecionada para avaliação manual. O procedimento de avaliação foi o mesmo das etapas anteriores, com a classificação das relações em 4 categorias (a pontuação 3 corresponde a uma relação ótima, a pontuação 0 a uma relação errada), e os resultados estão na tabela 13:

[37] A maior quantidade de relações extraídas no corpus genérico também é um indicativo de que as regras podem ser aplicadas com sucesso não com o objetivo de criar ontologias, mas talvez como uma ferramenta de auxílio a lexicógrafos.

Classificação	Qtd de relações	Exemplos
3	664 (81.6%)	Andrade Gutierrez < empresas Flashdance < filmes
2	23 (2.8%)	Barata Ribeiro < ruas do bairro Camboja < países asiáticos e africanos
1	33 (4%)	Ciro Gomes < lideranças Bertrand Russell < visitantes
0	93 (11.4%)	Antônio Britto < PMDB Billie Holliday < século

Tabela 13: Resultados da avaliação de nomes próprios no corpus genérico

A quantidade de relações classificadas como 3 (relações corretas), 81.6%, corresponde ao maior índice de acertos encontrado neste trabalho, maior inclusive que os resultados obtidos no corpus saúde, que já havia passado por um filtro prévio para eliminar erros puramente sintáticos, como erros decorrentes da ambigüidade do sintagma preposicionado ou de orações encaixadas no sintagma. Ou seja, 81.6% de acertos referem-se à aplicação das regras no corpus bruto. É exatamente a aplicação no corpus bruto que levou a um número relativamente alto de relações classificadas como 0 (relações erradas). Os erros nessa classe se devem, em sua maioria, à ambigüidade do sintagma preposicionado. O quadro 14 mostra alguns exemplos de relações erradas e as frases de onde foram extraídas.

Relação extraída	Frase do corpus
Cream < rock	...bandas de rock como Cream, ...
Breckenridge < esqui	...freqüenta estações de esqui como Breckenridge,...
Banco Mundial < financiamento	...provêm de organismos internacionais de financiamento como Banco Mundial, ...

Quadro 14: relações extraídas de frases com ambigüidade no SPrep

É importante observar, contudo, que mesmo com a grande ambigüidade (e freqüência na língua) dessas estruturas, as regras HHiper e HHipo tiveram um ótimo desempenho, já que não apenas 81% das relações estava correta, mas também porque diversas estruturas com o SPrep foram corretamente extraídas, como mostra o quadro 15.

George Miller < fundadores da ciência cognitiva	Genebaldo Correa < depoentes da primeira fase da CPI
Che Guevara < personagens da revolução	Elvis Presley < roqueiros dos anos 50
Beth Carvalho < puxadores de sambas	Humphrey Bogart < atores do cinema

Quadro 15: Relações corretamente extraídas que contêm SPrep.

A análise das relações classificadas como 1 (relações muito gerais para serem úteis) revelou que 33% dos erros é decorrência de um fenômeno já observado na análise dos resultados das regras: substantivos hiperônimos que possuem uma natureza relacional, como ilustram (a) e (b).

(a) Coréia<vizinhos (b) Compaq<concorrentes

Os seguintes substantivos relacionais foram encontrados no corpus: *adversário*, *irmã*, *vizinho*, *amigo*, *concorrente*. Além destes, outros substantivos hiperônimos que também indicam sistematicamente a necessidade de um complemento, embora não expressem relações entre indivíduos, apareceram com freqüência: *fabricante*, *visitante*, *criador*.

A multiplicidade de sentidos dos nomes próprios, característica que deve ser levada em conta no momento de sua classificação semântica, também é explicitada com a metodologia, como mostram os exemplos (c), (d) e (e):

(c) Argentina
- países
- times

(d) Austrália
- ilhas do Pacífico
- lugares
- países

(e) Chico Buarque
- artistas
-músicos brasileiros
- personalidades
- cinquentões

Por fim, os resultados da classificação semântica de nomes próprios no corpus genérico sugerem que a aplicação das regras de hiperonímia pode ser uma aliada em sistemas de reconhecimento de entidades mencionadas. Categorias como *autores*, *locais*, *países*, *cidades*, *bairros*, *marcas*, *empresas*, *pessoas*, *gente*, *jogadores*, além de conterem uma grande quantidade de nomes próprios, obtiveram 100% de acerto (exceto a categoria *cidades*). Os quadros 16, 17 e 18 mostram os resultados de *empresas*, *autores* e *países*.

empresas:Brasif Comercial, Eterbrás, General Mix Import-Export, Gensen Corp, Life Extension Foundation, Love and Kisses, Soccer Beach Company, Viação Auri Tupi, Água de Cheiro, Alcoa, AM / PM, Andrade Gutierrez, Arbi, Banco Francês e Brasileiro, Banco Nacional, Banco Noroeste, Banco Real, Boeing, Boston de o Brasil, Brittish Petroleum, Caesar Park Hotel, Carrefour, Chrysler, Citibank, Citrovita, Coca-Cola, Coelho, Compton's Nem Media, Discis Knowledge Research, Dupont, Flytour, Ford, Glaxo, grupo Gerdau, Interpass Club, Itambé, Jacadi, Kurzweil Music Systems, Lloyds Bank, Moinho Santista, Montreal Informática, Nacional Seguros, Nestlé, Norrau Informática, Papel Simão, Parmalat, Pinguim, Pirelli, Rio-Sul, Rummler-Brache Group, Sanbra, Santa Celina Mineradora, Shell, Souza Cruz, Stella Barros Turismo, Telerj, Tintas Coral, Varig, Vicunha

Quadro 16: Resultados da categoria *empresas*

autores: Anderson, Ariosto, Baudelaire, Berthold Goldschmidt, Bloch, Boccaccio, C. Geertz, Cabrera Infante, Carlos Felipe Moisés, Céline, Charles Dickens, Charles Mussel White, Clarice Lispector, Cláudio Guillén, Cláudio Willer, Curte Mayfield, Dante, Emily Brontë, Flaubert, García Márques, Georg Lukács, Goldman, Gramsci, H. Lefèbvre, Hannah Arendt, Hemingway, Herman Melville, Homero, Jack London, Jacques-émile Blanche, Jane Austen, José Cardoso Pires, Julia Kristeva, Kafka, Korngold, Krenek, Llosa, Ludwig Tieck, Maiakóvski, Mário de Andrade, Mark Twain, Marx, Maud Mannoni, Milan Kundera, Milton, Novalis, Octave, Octavio Paz, Paul Morand, Rabelais, René Welleck, Rimbaud, Robert Johnson, Roberto Piva, Schlegel, Schulhoff, Shakespeare, Thompson, Ullman, Umberto Eco, Van Tieghem, Voltaire

Quadro 17: Resultados da categoria *autores*

países:África do Sul, Alemanha, Alemanha Ocidental, Angola, Argélia, Argentina, Austrália, Bélgica, Brasil, Canadá, Chile, China, Colômbia, Coréia, Costa do Marfim, Egito, El Salvador, Espanha, Estados Unidos, EUA, Europa, Finlândia, França, Grã Bretanha, Guiné, Holanda, Honduras, Hong Kong, Hungria, Indonésia, Inglaterra, Irã, Iraque, Israel, Itália, Japão, Líbia, Malásia, Marrocos, Martinica, México, Namíbia, Nepal, Nova Zelândia, países de o Leste Europeu, Paraguai, Peru, Polônia, Portugal, Reino Unido, Rússia, Senegal, Singapura, Suécia, Suíça, Taiwan, Tanzânia, Ucrânia, União Soviética, Uruguai, Vietnã, Zaire

Quadro 18: Resultado da categoria *países*

8
Conclusões

Apresentei aqui subsídios para a elaboração automática de ontologias específicas quanto ao domínio. Embora a metodologia, em si, não seja nova, pois a correlação entre relações de hiponímia e a ocorrência de determinados padrões léxico-sintáticos em textos foi sugerida por Hearst (1992), acredito que as principais contribuições deste trabalho estão

(i) na proposta de novos padrões para a identificação da hiperonímia;

(ii) na adaptação e refinamento dos padrões existentes para o português;

(iii) na indicação de que o cruzamento das informações extraídas com os padrões, gerando inferências (produzindo conhecimento), é um processo válido e produtivo, desde que seja realizado em um corpus de domínio;

(iv) na adoção de uma perspectiva relativista com relação ao significado, que tem como conseqüência principalmente a análise de relações semânticas pouco convencionais, que poderiam ser consideradas "erro". Uma perspectiva relativista se mostra produtiva na medida em que legitima os dados vindos do corpus e as relações de significado que nele aparecem.

Com relação aos itens (i) e (ii), os padrões "tipos de" e "chamado" apresentaram um alto índice de precisão, embora tenham identificado poucas relações. A análise cuidadosa da estrutura "tais como" levou à identificação da estrutura variante "como", de alta freqüência no corpus, e a ajustes nas regras relacionados à presença de vírgula nas expressões. Uma análise minuciosa dos resultados iniciais dos padrões levou à criação da regra HHiper/ HHipo, que considera como sintagma hiperônimo / hipônimo apenas o último substantivo em SNs que contém sintagmas preposicionados – estruturas sintáticas altamente ambíguas na língua. Com isso, os resultados obtidos na extração foram muito positivos, principalmente se comparados aos obtidos em outros estudos (Hearst,

1998; Cederberg e Widdows, 2003). Porém, como já dissemos antes, a comparação deve ser vista com cautela, pois tanto a forma de avaliação – julgamento humano – é subjetiva, quanto as condições em que os trabalhos foram realizados foram diferentes (número de relações avaliadas, técnica de identificação das relações). Além disso, é preciso considerar que boa parte do sucesso na identificação é dependente de um fator "externo" – a etiquetagem de classes de palavras e de sintagmas nominais. Neste trabalho, o corpus etiquetado passou por uma revisão manual, na tentativa de minimizar a interferência de outras variáveis na identificação das relações, principalmente porque a estrutura do SN em português é mais complexa (tendo em vista a identificação automática) do que a do inglês (Oliveira e Santos, 2005). Ainda assim, os resultados da comparação servem como ilustração do potencial das regras.

Por fim, em favor da regras apresentadas aqui, lembro que, no padrão "como/tais como", 29% dos erros foi decorrente da presença de uma oração no sintagma hiperônimo / hipônimo[38], e que o extrator automático de sintagmas nominais subjacente à identificação das estruturas não reconhece SNs com orações. Conseqüentemente, é razoável supor que os resultados poderiam ser ainda melhores utilizando um modelo de SN que admita a identificação automática de orações.

Já no caso da regra "e outros", como 20% dos erros é decorrência de uma estratégia discursiva em que o hiperônimo retoma apenas o último elemento sa coordenação, uma forma de melhorar a precisão seria ajustar a regra para considerar apenas o último substantivo. Nesse caso, embora haja alguma perda na abrangência, a maior precisão pode ser útil, por exemplo, para uma etiquetagem de corpus de treino para sistemas de aprendizagem automática.

Os resultados do padrão "conhecido/a/os/as", que possibilitaria a inclusão de relações de co-referência na ontologia, foram desanimadores, pois apresentaram uma grande ambigüidade entre a expressão de co-referência e de hiperonímia. Em experimentos-piloto, não descritos neste trabalho, foram testadas também a identificação automática de apostos[39] e de orações explicativas[40] – construções

[38] Em "*fatores de risco como o hábito de fumar...*" é extraída a relação *fatores de risco* > *hábito*

[39] Exemplos de aposto:

(a) [*Metoprene, substância análoga a o hormônio juvenil de os insetos*,] que atua em as formas imaturas (larvas e pupas) , impedindo...

interessantes por também expressarem relações de co-referência. Porém, os resultados da identificação automática foram decepcionantes, o que levou à exclusão destas estruturas da metodologia. É importante salientar, contudo, que o problema não foi de ambigüidade das estruturas, como no caso do padrão "conhecido/a/os/as como", mas de natureza computacional: a identificação automática foi ineficaz. As estruturas são boas candidatas à expressão de co-referência, e merecem uma investigação detalhada quanto à possibilidade de identificação automática.

Com a exclusão dessas estruturas, que ofereceriam à ontologia relações de co-referência, a ontologia ficou apenas com as relações de hiperonímia, nisto se assemelhando a taxonomias.

Os resultados demonstraram, também, que freqüentemente nem todas as relações possíveis serão explicitadas na ontologia, indicando a necessidade de um trabalho humano complementar. Não há, por exemplo, nos resultados, uma relação entre a taxonomia de *animais* e a taxonomia de *mamíferos*. Isto nos faz ver com alguma cautela a afirmação de que "as categorias emergem do corpus" – sim, emergem, mas relações relevantes podem não emergir. Por outro lado, em uma visão otimista, é possível imaginar que em um corpus maior o problema seja minimizado.

A construção automática de ontologias a partir de grandes corpora é interessante tanto por reduzir a preocupação com o conhecimento a ser codificado, visto que esse conhecimento estaria no corpus, quanto por permitir a automação do processo, facilitando o trabalho de atualização. O que se tem, ao final, é um deslocamento do problema: em certa medida, passa-se para o corpus a "responsabilidade" de direcionar a construção da ontologia.

Investigações sobre a forma de avaliação de ontologias construídas automaticamente a partir de corpus são de fundamental importância, mas ainda não atingiram resultados satisfatórios. A versão simplificada da proposta de

---

(b) Estudos realizados em algumas áreas endêmicas de o estado de São Paulo utilizando a reação de imunofluorescência indireta, em comparação ao [*exame parasitológico de fezes, Kato-Katz,*] mostraram ..

[40]Exemplos de orações explicativas:

(a) Atualmente , a resistência à [*cloroquina, que é o antimalárico mais barato e mais amplamente usado*,] é comum em a África.

(b) ...foram devidas às [*doenças cardiovasculares, que são a primeira causa de morte em todas as grandes regiões de o país*,] com mortalidade proporcional...

avaliação de Brewster et al. (2004), que sugere uma comparação entre os termos relevantes presentes no corpus e os termos da ontologia, parece viável , justamente por prescindir de Wordnet e de um corpus semanticamente anotado. Porém, embora a metodologia verifique a adequação entre corpus e ontologia, não há como assegurar a correção das relações semânticas entre os termos. A proposta de Etzioni et al. (2005), de validação das relações por meio de busca por determinadas expressões na Web ("X é um Y"), pode ser um bom complemento nesse sentido. O principal problema desta abordagem é que, para a língua portuguesa, mecanismos de busca como o Google desconsideram acentos, o que leva a resultados indesejados.

A forma de avaliação utilizada aqui – validação manual – embora útil por permitir alguma comparação com outros trabalhos, é falha principalmente por não oferecer uma medida confiável nesta comparação. Julgamentos humanos são subjetivos, e um dos motivos para se sustentar a informação da ontologia em corpus é justamente a tentativa de minimizar esta subjetividade.

Retomando os critérios a que ontologias devem atender segundo Brewster e Wilks (2004) percebemos que todos foram atendidos, exceto o critério 5, que trata da origem dos dados para a construção da ontologia (documentos e uma taxonomia já existente), por razoes óbvias.

O critério 1, coerência interna, é atendido uma vez que as relações são extraídas de um corpus específico do domínio e é razoável supor que, em um mesmo domínio, haja coerência entre os usos dos termos. O critério 2, herança múltipla, também foi atendido, já que um mesmo termo pode ter mais de um pai na ontologia. Como os algoritmos de extração são simples, imagino que não haja complexidade na computação, o que está de acordo com o critério 3. Por fim, como os rótulos das categorias são os próprios termos extraídos, o critério 4, que aponta para a necessidade de nós com rótulos únicos, e não com rótulos que são grupos de palavras, também está atendido.

Um último comentário com relação aos resultados diz respeito aos nomes próprios. Embora o objetivo inicial do trabalho não tenha sido a classificação semântica de nomes próprios, tarefa que pertence à área de Reconhecimento de Entidades Mencionadas (REM) (ou NER – Named Entity Recognition, subárea da Extração de Informação), quando a metodologia foi aplicada a um corpus geral, composto por notícias de jornal, o grande número de relações envolvendo essas

estruturas mostrou que as regras podem ser uma ótima ferramenta para a extração de entidades mencionadas. Lembro novamente, contudo, que o corpus passou por uma revisão manual, o que minimizou consideravelmente a quantidade de erros decorrentes de dificuldades no processo de segmentação (anterior ao processo de classificação semântica). Assumindo, novamente em uma visão otimista, que a tarefa de segmentação de nomes próprios já esteja resolvida, persistem outros problemas relativos à natureza gramatical da categoria, e que irão interferir em sua classificação semântica: *AIDS* é um nome próprio? Em caso afirmativo, subentende-se, portanto, que um critério para uma palavra ser considerada nome próprio é constituir uma sigla (pois o que mais difere *AIDS* de *sarampo*, *gripe* etc?) Mas até que ponto *AIDS* é ainda reconhecida como sigla, e não como palavra simples da língua (vide *aidético*)? E *doença de Chagas*, *Mal de Alzheimer* ? Também são nomes próprios?

Por fim, lembro que a metodologia se beneficiaria com a identificação de expressões multi-vocabulares (EMVs) nominais no corpus. Embora os critérios de identificação de EMVs sejam controversos (Oliveira et al. 2004), a percepção de que determinadas combinações nominais, principalmente as de estrutura *Substantivo + Preposição "de" + Substantivo*[41] devem ser consideradas um único item lexical tem implicações importantes sobretudo na aplicação das regras HHiper / HHipo. O fato de EMVs nominais poderem ser identificadas com sucesso por meio de testes estatísticos, já que suas estruturas são, muitas vezes, sintaticamente transparentes, torna a incorporação dessas estruturas viável a curto-prazo. A transparência sintática de EMVs nominais, porém, tem conseqüências na aplicação da regra HiperN. Em *dor de cabeça*, por exemplo, é interessante que a regra seja empregada, originando o hiperônimo *dor*. Já em *pé de atleta*, a criação do hiperônimo *pé* seria um problema. A aplicação, nas EMVs nominais, de uma medida de similaridade capaz de avaliar a transparência sintática dessas construções seria útil para a identificação de EMVs que não estariam sujeitas à aplicação da regra HiperN.

8.1. Desdobramentos

Embora o objetivo inicial da ontologia tenha sido auxiliar tarefas que envolvem o processamento automático de textos, os resultados mostraram que a metodologia também pode ser de grande valia para investigações lexicográficas e lingüísticas. Nesse sentido, o insucesso dos resultados das inferências no corpus genérico pode ser visto como conseqüência de um "efeito colateral" positivo, pois a aplicação das regras no corpus possibilitou dois importantes achados: um tratamento para a classificação semântica de nomes próprios e um auxílio para lexicógrafos na tarefa de elaboração de dicionários.

8.1.1. Desdobramentos "mais" lingüísticos

De um ponto de vista lexicográfico, as relações entre os termos podem ser uma fonte valiosa para a observação dos contextos de ocorrência das palavras, contribuindo para a elaboração de dicionários e de léxicos específicos. A análise do comportamento das palavras ajuda na identificação dos seus múltiplos usos, fornecendo material para um processo preciso, empiricamente motivado e objetivo de atribuição de sentido.

Outro trabalho interessante relacionado à descrição do português é a caracterização formal, para posterior identificação automática, dos substantivos *relacionais*, aqueles que expressam relações entre indivíduos. como *pai*, *amigo*, *vizinho*, *adversário*, *concorrente*, *fundador*, *membro*, etc. A tarefa de classificação semântica de nomes próprios também se beneficiaria bastante deste tipo de informação.

A elaboração de critérios formais para a identificação automática de "adjetivos gerais", nos moldes da proposta de Oliveira (2006) de caracterização do substantivo-suporte, também seria de grande valia para tarefas de PLN.

[41] Alguns exemplos retirados do corpus: *prisão de ventre*, *atestado de óbito*, *taxa de natalidade*, *taxa de mortalidade*, *dor de cabeça*, *cinto de segurança*.

8.1.2.
Desdobramentos "mais" computacionais

Do ponto de vista do PLN, um trabalho interessante é aplicar técnicas de clusterização para distinguir grupos de palavras similares, utilizando como *seed words* palavras que já estão na ontologia, e verificar se o hiperônimo das *seed words* pode ser também hiperônimo das palavras do *cluster*. Com isso, haveria um aumento significativo da ontologia, com o acréscimo de co-hipônimos.

Outra possibilidade de trabalho é explorar de forma mais sistemática as técnicas de extração de informação na elaboração de ontologias. Por exemplo: excetuando-se os verbos auxiliares, os verbos mais freqüentes no corpus de saúde são *causar* e *evitar*. Supõe-se, portanto, que tais verbos expressem relações relevantes para o domínio de saúde. Em seguida, deve ser possível identificar, automaticamente, os sujeitos e objetos dos verbos, isto é, *X causa Y* e *X evita Y*. Desse modo, criam-se, semi-automaticamente, *templates* para a extração de mais informações.

Com relação aos padrões léxico-sintáticos utilizados neste trabalho, que podem ser considerados padrões de *templates* de EI, a principal vantagem está na generalidade: são padrões que podem ser aplicados a qualquer domínio, a qualquer tipo de texto – e o mesmo se aplica aos padrões referentes ao aposto e orações explicativas, não implementados.

8.2.
Considerações finais

Os resultados positivos da metodologia, tanto relativos ao corpus de domínio como ao corpus geral, indicam que sua aplicação pode ser uma importante aliada na elaboração de ontologias. Os resultados são decorrentes de uma análise lingüisticamente motivada e podem – devem – ser complementados com estratégias computacionais.

Uma estratégia utilizada, mas pouco vista em trabalhos de PLN, é a análise sistemática dos erros. Embora esta seja, sem dúvida, uma tarefa penosa, é de extrema valia para um entendimento de "*por que as coisas não estão acontecendo como o esperado*", principalmente quando estamos tratando de língua (em oposição a números). A elaboração das regras HHiper/HHipo, por exemplo, foi

decorrente de análise dos erros. A avaliação dos resultados – e dos erros – em termos das tradicionais medidas de precisão e abrangência não fornece pistas para aquilo que só a observação humana é capaz de descobrir, pois informam "apenas" o quanto os resultados obtidos ficaram distantes do ideal.

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. Suas principais desvantagens são a alta dependência de um corpus etiquetado e a dificuldade de avaliação sistemática (e de comparação) dos resultados.

9
Referências bibliográficas

AIRES, R.V.X.; ALUÍSIO, S.M. Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente. Série de Relatórios do NILC, NILC-TR-01-8, 2001.

ARISTÓTELES - Coleção Os Pensadores, São Paulo: Nova Cultural, 1991

BACKER, G. e HACKER, P. An analytical Commentary on Wittgenstein's Philosophical Investigations. Volume 1. Oxford: Blackwell, 1984.

BASILIO, M.. Introdução: questões clássicas e recentes na delimitação de unidades lexicais. In: M. BASILIO (org.) A Delimitação de Unidades Lexicais. PaLavra 5, Volume Temático I. Rio de Janeiro: PUC-RIO, p. 9-18, 1999.

BASILIO, M.; OLIVEIRA, C.; E GARRÃO, M.. A não-delimitação de unidades lexicais. In: HENRIQUES, C. (org) , Linguagem Conhecimento e Aplicação: estudos de língua e lingüística. Editora Europa, 2003.

BECHARA, E.. Moderna Gramática Portuguesa. 37ª ed. Rio de Janeiro, Lucerna, 1999.

BIBER, D.; CONRAD, S.; E REPPEN, R.. Corpus Linguistics: Investigating Language Structure and Use. Cambridge, UK: Cambridge University Press. 1998.

BOGURAEV, B.; PUSTEJOVSKY, J. Issues in Text-based Lexical Acquisition. In: BOGURAEV, B. & PUSTEJOVSKY, J (orgs.). Corpus Processing for Lexical Acquistion. Cambridge, Massachusetts: MIT Press. 1996

BREWSTER, C. e WILKS, Y. Ontologies, Taxonomies, Thesauri: Learning from Texts. In: Proceedings The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop, Kings College, London, UK, 2004.

BREWSTER, C., ALANI, H., DASMAHAPATRA, S. e WILKS, Y. Data Driven Ontology Evaluation. In: Proceedings of International Conference on Language Resources and Evaluation(LREC 2004), Lisboa, Portugal, 2004.

BREWSTER, C., CIRAVEGNA, F. e WILKS, Y. Knowledge Acquisition for Knowledge Management: Position Paper. In: IJCAI-2001 Workshop on Ontology Learning, Seattle, USA, 2001.

BREWSTER, C., IRIA, J., CIRAVEGNA, F., WILKS, Y. The Ontology: Chimaera or Pegasus. Dagstuhl workshop on Learning for the Semantic Web, Dagstuhl, Germany, 2005.

BUITELAAR, P. Semantic lexicons: between ontology and terminology. In: Proceedings of Ontolex: Ontologies and Lexical Knowledge Bases. 2000. OntoText Lab. Sofia, Bulgaria, 2001.

CARABALLO, S. Automatic construction of a hypernym-labeled noun hierarchy from text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), 120-126, 1999.

CEDERBERG, S. e WIDDOWS, D. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: Seventh Conference on Computational Natural Language Learning (CoNLL-2003), Edmonton, Canada, 111-118, 2003.

CONDAMINES, A. e REBEYROLLE, J. Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In L'HOMME, M.-C., JACQUEMIN, C. e BOURIGAULT, D. (eds.), Recent Advances in Computational Terminology, Amsterdam/Philadelphia: John Benjamins Publishing Company, p.127-148, 2000.

CRUSE, D. Lexical Semantics. Cambridge, Inglaterra: Cambridge University Press,1986.

CRUSE, D. Meaning in Language: An Introduction to Semantics and Pragmatics. UK: Oxford University Press. 2004.

DIAS-DA-SILVA, B. Wordnet.Br: An exercise of human language technology research. In: Revista PaLavra, no. 12, 2004. Série Linguagem. Volume Temático: Processamento Automático do Português. Org.: DIAS, M. C. e QUENTAL, V., Edições Galo Branco, p. 15-24, 2004.

DIAS-DA-SILVA, B., DI FELIPPO, A. e HASEGAWA, R. Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations. In: VIEIRA, R., QUARESMA, P., VOLPES NUNES, M.G., MAMEDE, N. OLIVEIRA, C. e DIAS, M.C. (eds.), 7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006), Springer, pp. 120-130, 2006.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., POPESCU, A., SHAKED, T., SODERLAND, S., WELD, D. S., e YATES, A. Unsupervised named-entity extraction from the web: an experimental study. Artificial Intelligence. 165, 1, 91-134, 2005.

FELLBAUM, C. WordNet: An Electronic Lexical Database, MIT Press, 1998.

FREITAS, M.C; GARRÃO, M.; OLIVEIRA, C.; SANTOS, C. N. e SILVEIRA, M.C. 2005. A anotação de um corpus para o aprendizado supervisionado de um modelo de SN. In: Anais do XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, São Leopoldo, 2005.

GARRÃO, M. O corpus não mente jamais : sobre a identificação e uso de combinações multivocabulares do tipo verbo mais sintagma nominal; Tese de Doutorado – Rio de Janeiro : PUC, Departamento de Letras, 2006.

GLOCK, H.J. Dicionário Wittgenstein. Rio de Janeiro: J. Zahar, 1997.

GRUBER, T.. Toward principles for the design of ontologies used for knowledge sharing. Int. Journal of Human-Computer Studies, v. 43, p.907-928, 1993.

GRUBER, T. What is an ontology?.1996. Disponível em: http://www-ksl.stanford.edu/ kst/ what-is-an-ontology.html. Acesso em: 25 nov. 2006.

GUARINO, N. Understanding, building and using ontologies. In: PROCEEDINGS OF KNOWLEDGE ACQUISITION FOR KNOWLEDGE-BASED SYSTEMS WORKSHOP. 10. 1996. Disponível em: <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/ guarino.html# Heading4>. Acesso em: 25 nov. 2006

HEARST, M. Automated discovery of WordNet relations. In: Fellbaum, Christiane, ed., WordNet: An Electronic Lexical Database, MIT Press, May 1998.
HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the the 14th International Conference on Computational Linguistics, Nantes, 1992.

HOVY, E. Comparing Sets of Semantic Relations in Ontologies. In: Green, R., Bean, C. e Myaeng, S. editors, The Semantics of Relationships: An Interdisciplinary perspective, Kluwer, p. 91-110, 2002.

HOVY, E. Methodologies for the Reliable Construction of Ontological Knowledge. In: F. Dau, M.-L. Mugnier, and G. Stumme (eds), Conceptual Structures: Common Semantics for Sharing Knowledge. Proceedings of the 13th Annual International Conference on Conceptual Structures (ICCS 2005). Kassel, Germany. Springer Lecture Notes in AI 3596, pp 91–106, 2005.

ILARI, R. e GERALDI, J. W. Semântica. São Paulo: Ática, 1985.

JACQUEMIN, C. Syntagmatic and paradigmatic representations of term variation. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), University of Maryland, p. 341-348, 1999.

JACQUEMIN, C., DAILLE, B., ROYANTÉ, J., e POLANCO, X.. In vitro evaluation of a program for machine-aided indexing. Inf. Process. Manage. 38, 6, p. 765-792, 2002.

KALFOGLOU, Y. E SCHORLEMMER, M. Ontology Mapping: The State of the Art. The Knowledge Engineering Review Journal, vol. 18:1, 1-31. Cambridge University Press, 2003.

KILGARRIFF, A. I Don't Believe in Word Senses. Computers and the Humanities, 31 (2), p.91-113, 1997.

KILGARRIFF, A. Thesauruses for Natural Language Processing .Proceedings of NLP-KE, Beijing, China, p.5-1, 2003.

LIN, D. e PANTEL, P. Concept discovery from text. In: Proceedings of the 19th international Conference on Computational Linguistics - Volume 1 (Taipei, Taiwan, August 24 - September 01, 2002). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-7, 2002.

LOBATO, L.. Adjetivo: Tipologia e interpretação semântica. Boletim da ABRALIN 14. 1993.

LYONS, J. Semântica. Martins Fontes, 1980.

MAEDCHE, A. e STAAB, S. Measuring Similarity between Ontologies. In: Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Madrid, Spain, LNCS/LNAI 2473, Springer, pp. 251-263, 2002.

MAEDCHE, A. e STAAB. 2000. Discovering conceptual relations from text. In ECAI-2000 - European Conference on Artificial Intelligence. Proceedings of the 13th European Conference on Artificial Intelligence. IOS Press, Amsterdam, p.321-325, 2000.

MANI, I. e MacMILLAN, R.. Identifying Unknown Proper Names in Newswire text. In: BOGURAEV e PUSTEJOVSKY (1996). Corpus Processing for Lexical Acquisition. Oxford University Press, 1996.

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

MARCONDES, D. Pragmática. Rio de Janeiro: Jorge Zahar, 2005.

MARQUES, M. H. D.. Léxico de alta freqüência na língua portuguesa. In: HEYE, J. (org). Flores verbais, uma homenagem lingüística e literária para Eneida do Rego Monteiro Bomfim no seu 70° aniversário. Rio de Janeiro: 34 Editora, 1995. p. 247-282, 1995.

MARTINS, H. Metáfora e Polissemia no estudo das línguas do mundo: uma aproximação não representacionista. Tese de Doutorado inédita, UFRJ, 1999.

MARTINS, H. Três Caminhos na Filosofia da Linguagem. In MUSSALIM, F; BENTES, A.C. (orgs.). Introdução à Lingüística. Volume III, São Paulo: Cortez Editora, 2004. p. 439-474.

McDONALD, D.. Internal and external evidence in the identification and semantic categorization of proper names. In: BOGURAEV e PUSTEJOVSKY (1996). Corpus Processing for Lexical Acquisition. Oxford University Press, 1996.

MORIN, E. e JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. Computer and the Humanities, vol. 38 (4), 343-362l, 2004.

NIRENBURG, S. e WILKS, Y. What's in a symbol: Ontology, representation, and language. Journal of Experimental and Theoretical Artificial Intelligence, 13(1):9–23, 2001

NOY, F. N.; GUINNESS, D. L. Ontology development 101: a guide to create your first ontology. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, Stanford University, Stanford, CA.2001. Disponível em: <http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.doc>. Acesso em: 26 nov 2006.

OLIVEIRA, C. M. e FREITAS, M. C. Classes de palavras e etiquetagem na Lingüística Computacional. Artigo Submetido, 2006.

OLIVEIRA, C. M. G. M. e SANTOS, C. N. . Extracting Brazilian Portuguese Noun Phrases from Texts with TBL. Archives of Control Sciences, Varsóvia, v. 15(LI), n. 3, p. 251-262, 2005.

OLIVEIRA, C.M.. O Substantivo-suporte:Critérios Operacionais de Caracterização. Rio de Janeiro, 2006. 116p. Tese de Doutorado — Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro, 2006.

OLIVEIRA, C., FREITAS, M.C., GARRÃO, M., NOGUEIRA, C. e ARANHA, C.. A extração de expressões multi-vocabulares: uma abordagem estatística. In: Revista PaLavra, no. 12, 2004. Série Linguagem. Volume Temático: Processamento Automático do Português. Org.: DIAS, M. C. e QUENTAL, V., Edições Galo Branco, p. 172-192, 2004.

PATRICK, P. e RAVICHANDRAN, D. Automatically Labeling Semantic Classes. In: Proceedings of Human Language Technology / North American chapter of the Association for Computational Linguistics (HLT/NAACL-04). p. 321-328, 2004.

PEREIRA, M.T. Palavras denotativas: temas e problemas. In: HEYE, J. (org). Flores verbais, uma homenagem lingüística e literária para Eneida do Rego Monteiro Bomfim no seu 70° aniversário. Rio de Janeiro: 34 Editora, 1995. pp. 15-21, 1995.

PHILLIPS, W. e RILOFF, E. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In Proceedings of the Acl-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 Annual Meeting of the ACL, 125-132, 2002.

RILOFF, E. e SHEPHERD, J. A corpus-based approach for building semantic lexicons. In: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97), 117-124, 1997.

SAEED, J., Semantics. Blackwell Publishers. 1997.

SANTOS, C.N., OLIVEIRA, C.: Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: Anais do XXV Congresso da Sociedade Brasileira de Computação, Brasil, 2005.

SANTOS, Diana. "Introdução ao processamento de linguagem natural através das aplicações", in Elisabete Ranchhod (ed.), *Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações*, Lisboa: Caminho, p.229-259, 2001

SCHÜTZE, H. Automatic word sense discrimination. Computational Linguistics, 24(1):97-124, 1998.

SMITH, B. 2003. Ontology, In: FLORIDI, L. (ed.), Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell, 2003. p. 155–166.

SMITH, B. Ontology and Information Systems. Disponível em ontology.buffalo.edu/ontology(PIC).pdf. Acessado em 22/11/2006.

SNOW, R., JURAFSKY, D., e NG, A. Y. Learning syntactic patterns for automatic hypernym discovery, Advances in Neural Information Processing Systems 17, 2005.

SOWA, J. F. Building, sharing and merging ontologies. Tutorial. [S. 1. : s. n.], 1999. Disponível em: <http://users.bestweb.net/~sowa/ontology/ontoshar.htm>. Acesso em: 26 nov 2006.

SPARCK-JONES, K. Natural language processing: a historical review. Disponível em http://www.cl.cam.ac.uk/~ksj21/histdw4.pdf. Acessado em 20/11/206

TAYLOR, T. Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretation (Post-Contemporary Interventions). Duke University Press, 1992.

VELARDI, P., NAVIGLI, R., CUCHIARELLI, A., NERI, F. Evaluation of Ontolearn, a Methodology for Automatic Population of Domain Ontologies. In: Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, 2005.

VIEGAS, E., MAHESH, K., NIRENBURG, S., e BEALE, S. Semantics in Action. In: SAINT-DIZIER, P. (Ed), Predicative Forms in Natural Language and in Lexical Knowledge Bases. Dordrecht-Boston:Kluwer, 171-203, 1999. Disponível em http://ilit.umbc.edu/SergeiPub/SemantInAction98.pdf. Acessado em 19/12/2006.

VOSSEN, P. (Ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht, 1998.

VOSSEN, P. Ontologies. In: MITKOV, R. The Oxford handbook of computational linguistics. Oxford: Oxford University Press (2003)

WIDDOWS, D. e DOROW, B. A Graph Model for Unsupervised Lexical Acquisition. 19th International Conference on Computational Linguistics, Taipei. 1093-1099, 2002.

WIDDOWS. D. Unsupervised methods for developing taxonomies using syntactic and statistical information. In: Proceedings of HLT/NAACL 2003, Edmonton, Canada, 276-283, 2003.

WILKS, Y. IR and AI: traditions of representation and anti-representation in information processing. In: Proceedings of IEE Conference on IR and AI, Glasgow, 1999.

WILKS, Y. Ontotherapy: or how to stop worrying about what there is. Invited presentation, Ontolex 2002, Workshop on Ontologies and Lexical Knowledge Bases, 27th May. Held in conjunction with the Third International Conference on Language Resources and Evaluation - LREC02, 29-31 May, Las Palmas, Canary Islands, 2002.

WITTGENSTEIN, L. Investigações Filosóficas. Coleção Os Pensadores, São Paulo: Abril Cultural, 1979.

ZÚÑIGA, G. Ontology: Its transformation from philosophy to information systems. Proceedings of the Second International Conference (FOIS '01), New York: ACM Press, 187-197, 2001

10
Anexos

ANEXO 1: 100 adjetivos mais freqüentes no corpus saúde

1. maior
2. social
3. primeiro
4. grande
5. importante
6. novo
7. público
8. menor
9. baixo
10. alto
11. possível
12. humano
13. principal
14. médico
15. físicoo
16. necessário
17. presente
18. variável
19. específico
20. crônico
21. sexual
22. último
23. segundo
24. clínico
25. brasileiro
26. pequeno
27. melhor
28. responsável
29. hospitalar
30. grave
31. comum
32. básico
33. materno
34. diverso
35. positivo
36. significativo
37. relativo
38. semelhante
39. geral
40. próprio
41. superior
42. técnico
43. familiar
44. freqüente
45. político
46. individual
47. nacional
48. epidemiológico
49. seguinte
50. capaz
51. jovem
52. anterior
53. etária
54. disponível
55. local
56. fundamental
57. urbano
58. diferente
59. total
60. genético
61. infantil
62. econômico
63. existencial
64. feminino
65. municipal
66. masculino
67. inferior
68. sócio-econômico
69. rural
70. especial
71. mental
72. terceiro
73. alimentar
74. difícil
75. privado
76. pré-natal
77. precoce
78. normal
79. simples
80. populacional
81. terapêutico
82. adulto
83. bucal
84. direto
85. financeira
86. saudável
87. neonatal
88. amplo
89. inicial
90. central
91. profissional
92. múltiplo
93. nutricional
94. externo
95. preciso
96. complementar
97. cultural
98. forte
99. biológico
100. ambulatorial

ANEXO 2: Taxonomia de *infecções* (corpus saúde)

INFECÇÕES
—bronquite
—diarréia
—hepatites
—infecção aguda
— —meningococcemia fulminante
—infecções bacterianas
— —cólera
— —difteria
— —hanseníase
— —leptospirose
— —meningite
— —tuberculose
—infecções cutâneas
— —espinhas
— —furúnculos
— —machucados
—infecções não intestinais
— —infecção de o trato urinário
— —otite média
— —pneumonia
— —sepsis
—infecções raras em adultos
— —sarampo
— —varíola
—infecções respiratórias agudas
—infecções virais sistêmicas
— —herpes disseminado
— —sarampo
—infecções virais
— —HIV
— —sarampo
—infecções virais
— —rubéola
—pneumonia
—sarampo

ANEXO 3: Taxonomia de *agravos* (corpus saúde)

AGRAVOS -agravos à saúde --doenças ---acidente vascular cerebral ---Aids ---Alzheimer ---asma ---câncer ----câncer nasofaringeo ----carcinoma in situ ----linfoma de Burkitt ---cardiopatia ---cardiopatias ----cardiopatias isquêmicas de o coração -----doenças cerebrovasculares -----insuficiência cardíaca congestiva ----hipertensão ---cólera ---dengue ----dengue clássica ----dengue hemorrágica ---derrame cerebral ---diabete ---diabetes ---diarréias ---dislipidemia ---doença atópicas ----asma ----eczema ----febre do feno ----urticária ---doença crônica ----diabetes ----hipertensão ---doença mental ----esquizofrenia ---doenças auto-imunes ----artrite ---doenças cardiovasculares ----acidente vascular cerebral ----doenças isquêmicas de o coração ----insuficiência cardíaca ---doenças congênitas ----fenilcetonúria ---Doenças Coronarianas e Cerebrovasculares ---doenças crônicas severas ----diabete tipo 2 ---doenças crônico-degenerativas ----doenças cardiovasculares -----acidente vascular cerebral -----doenças isquêmicas de o coração -----insuficiência cardíaca	---doenças de natureza auto-imune ----lupus eritematoso sistêmico ---doenças de órgãos ----doenças imunológicas ----erros inatos de metabolismo ----malignidades ---doenças de transmissão respiratória ----Síndrome Respiratória Aguda Grave ----tuberculose ----varicela ---doenças endócrinas e auto-imunes ---doenças febris ----gripe ----hepatite ----malária ----sarampo ---doenças infecciosas intestinais ----cólera ---doenças infecciosas ----HIV ----sarampo ---doenças invasivas ----meningites ---doenças normalmente díspares ----diabetes ----HIV / AIDS ---doenças respiratórias crônicas ----asma brônquica ---doenças respiratórias ----bronquite ----pneumonia ---doenças sexualmente transmissíveis ----AIDS ---doenças tipicamente relacionadas com o lixo ----diarréias ----doenças de pele ----leptospirose ----parasitoses ---doenças transmissíveis ----HIV ---doenças transmitidas ----malária ---doenças vasculares ----acidentes vasculares cerebrais	----cardiopatia ---erlichiose ---febre amarela ---febre de as trincheiras ---febre maculosa ---febre Q ---gripes ---hanseníase ---hantavirose ---hepatite ---hepatites ---Hipertensão Arterial ---hipertensão ---HIV ---infecções crônicas ---infecções de o trato respiratório superior ---insuficiência renal ---leishmaniose ---leptospirose ---leucemia ---mal de Alzheimer ---mal de Parkinson ---malária ---mononucleose ---obesidade ---osteoporose ---parasitoses ---Parkinson ---pneumonia ---problemas dentários ---protozooses ---psicoses ---raiva ---resfriados ---sarna ---sífilis ---tifo ---toxoplasmose ---tuberculose -agravos relacionados a o trabalho --conjuntivite --dermatites --distúrbios ósteo-musculares --estresse --hepatite --perda auditiva --problemas oculares --varizes

ANEXO 4: Taxonomia de *complicações* (corpus saúde)

COMPLICAÇÕES
—anemia hemolítica
—cegueira
—choque
—complicações clínicas
— —embolismo pulmonar
— —hematúria
— —hipostenúria
— —infarto esplênico
—complicações neurológicas
— —cerebelite
— —encefalite
—complicações obstétricas
— —amniorrexe prematura
—complicações pós—cirúrgicas
— —linfedema de o braço afetado
—complicações respiratórias
— —broncopneumonia
— —dispnéia
— —enfisema
— —ruptura de o diafragma
—complicações sistêmicas
— —abcesso hepático
— —disenteria
—comprometimento de o sistema nervoso central, de os rins e pulmões, de as lesões vasculares
—diarréia
—encefalite
—febre reumática
—hemorragia
—infarto de o miocárdio
—infecções respiratórias agudas
—infecções respiratórias
—meningite
—nefrite
—otite
—peritonite
—pneumonia bacteriana
—pneumonia
—rompimento de o baço
—sarampo
—septicemia
—trombocitopenia

ANEXO 5: Taxonomia de *produtos* (corpus genérico)

PRODUTOS
—açúcar
—algodão
—alimentos
— —açúcares
— —alimentos não-laxativos
— — —arroz
— — —banana-maçã
— — —batata
— — —maçã
— — —pera
— —batata-doce
— —batatas
— —carne
— —carne-seca
— —cascas
— —cenoura
— —cereais
— —farinha
— — —biscoitos
— — —confeitaria
— — —massas
— — —produtos de panificação
— —folhas
— —frutas secas
— — —damascos envolvidas em chocolate
— — —figos
— — —tâmaras
— —frutas
— — —abacaxi
— — — —Banespa
— — —acerola
— — —frutas brasileiras
— — — —acerola
— — — —cupuaçu
— — —frutas secas
— — — —damascos envolvidas em chocolate
— — — —figos
— — — —tâmaras
— — —granola
— — —mamão
— — —manga
— — —maracujá
— — —pedaços de banana
— — —tâmara
— —gordura
— —legumes
— — —Tomates verdes
— — —vagens
— —leite
— —maionese
— —mamão
— —manteiga
— —massas
— —melão
— —óleos
— —ovos
— —paio
— —pão
— —saladas
— —talos
— —toucinho
— —tubérculos
— —verduras
—bebidas
—biscoitos dinamarqueses
—biscoitos
—cacau
—café
—calçados
—carne
—carnes
—carpetes
—cereais matinais
—charutos
— —populares
—cobertores de campanha
—conservas de frutas
—corrosão
—cortinas
—creme de leite
—discos
—elixir de pedras
—equipamentos eletrônicos
—farinhas nobres
—feijão
—feltros
—fertilizantes
—forros
—frutas
— —abacaxi
— — —Banespa
— —acerola
— —frutas brasileiras
— — —acerola
— — —cupuaçu
— —frutas secas
— — —damascos envolvidas em chocolate
— — —figos
— — —tâmaras
— —granola
— —mamão
— —manga
— —maracujá
— —pedaços de banana
— —tâmara
—fumo
—gasolina
—geléias
—laranja
—leite condensado
—material de papelaria
—milheto
—milho
—óleo
—ovos
—pão
—passagens aéreas
—pesticidas
—pétalas de flores
—pisos
—plásticos
—rum
—sabonetes
—sal
—sisal
—soja
—produtos aquáticos
— —maiôs
— —toucas
—produtos de origem francesa
— —móveis
— — —carteiras
— — —pranchetas
— —tecidos
— — —algodão
— — —crepe
— — —gabardine
— — —tecidos nobres
— — — —microfibra
— — — —seda
— — — —veludo molhado
— — —tecidos sintéticos
— — — —ciré
— — — —jeans tratado com silicone
— — — —PVC
—produtos elaborados a partir de a madeira
— —celulose
— —papel
— — —Jacques Rigaut
—produtos ligados à Copa
— —camisetas
— —uniformes
—produtos ligados ao vôlei
— —bonés
— —camisetas
—produtos selecionados por Gil
— —discos
— —livros de arte
—produtos típicos
— —castanha de caju
— —linho
— —madeira
— — —hashis
— —redes
— — —Agua de Cheiro
— — —Localiza
— — —McDonald's
— — —Multicoisas
— — —Paes Mendonça
—queijo
—remédios
— —remédios mundialmente usados
— — —Dienpax
— — —Lexotan
— — —Rohypnol
—utilidades domésticas
— —panelas
— —jogo de travessas
—vestuário
—videocassetes
—vídeos
—vinhos
—sorgo
—tapetes
—arroz
—automóveis

ANEXO 6: Taxonomia de *utensílios* (corpus genérico)

UTENSÍLIOS
— utensílios contaminados
— — objetos
— — — armas
— — — conchas
— — — equipamentos
— — — — câmbio totalmente automático
— — — — churrasqueiras
— — — — cinto de segurança
— — — — controle de tração
— — — — drives de CD- ROM
— — — — equipamentos dos soldados
— — — — — cantis
— — — — — quepes
— — — — equipamentos odontológicos
— — — — — dentaduras
— — — — — tempo atrás brocas
— — — — equipamentos padrão SCSI
— — — — — acionador de CD- ROM
— — — — — Midi
— — — — — um gravador digital
— — — — equipamentos[acima]
— — — — estepe
— — — — extintor de incêndio
— — — — freio ABS
— — — — macaco
— — — — mesas
— — — — playground
— — — — programas
— — — — — aplicativos de uso pessoal
— — — — — circos
— — — — — educação
— — — — — — colégios
— — — — — — compra de material didático
— — — — — — cursos
— — — — — — — enfermagem
— — — — — — — engenharias civil
— — — — — — — fisioterapia
— — — — — — — obstetrícia
— — — — — — — odontologia
— — — — — entretenimento com jogos
— — — — — espetáculos de marionetes
— — — — — horários de outros museus
— — — — quadras poliesportivas
— — — — sauna
— — — — semeadoras
— — — — sismógrafos
— — — — suspensão ativa
— — — — triângulo
— — — esculturas
— — — ferramentas
— — — — ancinho
— — — — computadores
— — — — — impressoras
— — — — — modems
— — — — e- mail
— — — — hipertexto
— — — — intuição
— — — — linguagem Logo
— — — — multimídia
— — — — perturbação
— — — — software educacional
— — — luminárias
— — — móveis
— — — — carteiras
— — — — pranchetas
— — — objeto feminino
— — — — anéis
— — — — batons
— — — — brincos
— — — objeto material
— — — — objetiva
— — — objetos banais
— — — — bancos de madeira
— — — — cadeiras
— — — — leques
— — — objetos de o desejo de a delegação brasileira
— — — — equipamentos de som
— — — — microcomputadores
— — — — TVs
— — — objetos de uso cotidiano
— — — — flores artificiais
— — — — livros
— — — — pão
— — — — uma garrafa de água
— — — objetos de valor
— — — — dinheiro
— — — — jóias
— — — objetos pertenecentes às culturas
— — — — obras
— — — — — esculturas
— — — — — Lucíola
— — — — — móbiles
— — — — — Perfil de Mulher
— — — — — Senhora
— — — — vestimentas cerimoniais
— — — objetos pessoais
— — — — livros
— — — — quadros
— — — pedras
— — — penas
— — — — multas elevadas
— — — — prestação de serviços comunitários
— — — — suspensão de a habilitação
— — — povos
— — — — arianos
— — — — vândalos
— — — pratos
— — — — aceto balsâmico
— — — — costelinha com samambaia
— — — — ensopados
— — — — filetto tartufo
— — — — frango com quiabo
— — — — hossomakis
— — — — massas
— — — — scaloppine funghi porcini
— — — — sopas
— — — — sushis
— — — — tambaqui assado
— — — — tempuras a um preço fixo de R$ 18,00
— — — quadros
— — — radiorrelógios
— — — televisores
— — — vasos
— — — — linfáticos
— — — xícaras

Cont. Taxonomia de *utensílios*

UTENSÍLIOS
— *utensílios contaminados*
— — *objetos*
— — roupas
— — — bermudas
— — — coletes
— — — minissaias
— — — roupas casuais
— — — — botas tipo Timberland
— — — — jeans
— — — — t— shirts
— — — roupas masculinas
— — — — Intimo Due
— — — shorts
— — talheres

ANEXO 7: Taxonomia de *países* (corpus genérico)

PAÍSES
— África do Sul
— Alemanha Ocidental
— Alemanha
— Angola
— Argélia
— Argentina
— Austrália
— Balzac
— Bélgica
— Brasil
— Canadá
— Chile
— China
— Colômbia
— Coréia
— Costa do Marfim
— Egito
— El Salvador
— Espanha
— Estados Unidos
— EUA
— Europa
— Finlândia
— Formosa
— França
— Grã Bretanha
— Grécia
— Guiné
— Holanda
— Honduras
— Hong Kong
— Hugo
— Hungria
— Indonésia
— Inglaterra
— Irã
— Iraque
— Israel
— Itália
— Japão
— Líbia
— Malásia
— Marrocos
— Martinica
— México
— Namíbia
— Nepal
— Nova Zelândia
— país africano
— — Níger
— país continente
— — América do Norte
— país fictício
— — Moscou
— países africanos
— — Quênia
— — Senegal
— países asiáticos e africanos
— — Camboja
— — Cingapura
— — Etiópia
— — Laos
— países bilíngues
— — Canadá
— países de língua latina
— — Portugal
— países díspares
— — Alemanha
— — Brasil
— — Canadá
— — China
— — França
— — Itália
— — Mali
— países do Leste Europeu
— — Hungria
— — Polônia
— países islâmicos
— — Arábia Saudita
— países menos expressivos para o cristianismo
— — Chile
— — Equador
— — México
— Paraguai
— Peru
— Polônia
— Portugal
— Reino Unido
— República Autônoma da Tartária
— Rússia
— Senegal
— Singapura
— Sri Lanka
— Suécia
— Suíça
— Taiwan
— Tanzânia
— Ucrânia
— União Soviética
— Uruguai
— Vietnã
— Zaire

ANEXO 8: Taxonomia de *profissionais* (corpus genérico)

PROFISSIONAIS
— advogados
— — advogado hollywoodiano
— — — Alan Rothenburg
— — advogado nova-iorquino
— — — Leonard Finz
— babás
— Bob Wolfenson
— cientista político de o Iser
— Claudio Elizabetsky
— contadores
— dentistas
— economistas
— — Carlos Lessa
— — Fernando Henrique Cardoso
— — Francisco Weffort
— — Maria da Conceição Tavares
— editores de moda
— empregadas domésticas
— estilistas
— — Armani
— — Christian Lacroix
— — Dolce
— — Gabbana
— — Giorgio Armani
— — Inner Space
— — Jean-Paul Gaultier
— — Kenzo
— — Lagerfeld
— — Lolita Lempicka
— — Martin Margiela
— — Rei Kawakubo
— — Saint— Laurent
— — Thierry Mugler
— — Valentino
— — Versace
— — Vivienne Westwood
— faxineiros
— profissionais liberais
— — André Lara Resende
— — ex-negociador da dívida externa
— — sócio do Banco Matrix
— psicanalistas
— sociólogo
— motoristas
— nutricionistas
— políticos
— — Café Filho
— — Fleury
— — Lacerda
— — Luiz Inacio Lula
— — Mário Covas
— fotógrafos
— — Arthur Elgort
— — Edumund Collein
— — Erich Cosemuller
— — Eugen Batz
— — fotógrafos americanos
— — — Berenice Abbott
— — — Clarence John Laughlin
— — — Frederick Sommer
— — — Helen Levitt
— — — Horst P. Horst
— — fotógrafos brasileiros
— — — Annie Leibovitz
— — Gertrud Arndt
— — Gui Paganini
— — Herb Hitts
— — Herbert Bayer
— — Josef Albers
— — Kathina Both
— — Laszló Maholy-Nagy
— — Mario Sorrenti
— — Patrick Demarcheller
— — Steven Meisel
— — T. Luz Feininger
— — Warner Graef
— Gal Oppido
— ginastas
— janeiro
— jornalistas
— — Fernando Gabeira
— — Frei Beto
— — irregulares de o jornal Diário do Pará
— — irregulares
— — Scritta
— — Zuenir Ventura
— lava- pratos
— Leandro Piquet Carneiro
— médicos
— — médico francês
— — — Julian Offray de la Mettrie
— produtores
— — produtores californianos
— — — Chateau Montelena
— — — Duckhorn
— — — Firestone
— profissionais da área de saúde de todo o país
— — assistentes sociais
— — psicólogos
— — — psicólogos inovadores
— — — — Donald Broadbent
— — — — George Miller
— — — — Jerome Bruner
— — psiquiatras

ANEXO 9: Taxonomia de *conceitos* (corpus genérico)

CONCEITOS
— causalidade
— conceitos abstratos
— — leis econômicas
— — preceitos religiosos
— — princípios políticos
— — sistemas de crenças
— conceitos analíticos
— — esquemas
— — estratégias
— — estruturas
— — — estames
— — — folículos de Graaf
— — — nódulos linfáticos
— conceitos de trigonometria
— — cossenos
— — senos
— conceitos mais ou menos vagos
— conceitos molares
— — esquemas
— — estratégias
— — operações
— — — operações ilegais
— — — — contrabando
— — — — lavagem de dinheiro
— — — — tráfico de drogas
— — — operações similares
— — — operações similares
— conceitos vagos
— — exposição da marca
— — imagem
— — retorno
— conhecimento
— elaboração
— esquemas
— Gestão participativa
— habilidade
— imagens
— — cartoons
— — estrelas
— — — Al Hirt
— — — Allman Brothers
— — — Aretha Franklin
— — — B.B. King
— — — Beckenbauer
— — — Catherine Deneuve
— — — Chillan
— — — Denise Fraga
— — — estrelas americanas
— — — — Glenn Close
— — — — Jeremy Irons
— — — — Meryl Streep
— — — — Vanessa Redgrave
— — — — Winona Ryder
— — — estrelas do elenco rubro-negro
— — — — Romário
— — — — Sávio
— — — Etta James
— — — Gérard Depardieu
— — — Giulia Gam
— — — Glen Rice
— — — Grant Long
— — — Harold Miner
— — — Harrison Ford
— — — Hotel Zum Jaegerwirt
— — — José Wilker
— — — La Parva
— — — Larry Bird
— — — Little Richard
— — — Magic Johnson
— — — Marcello Mastroianni
— — — Michael Jordan
— — — Neville Brothers
— — — Pelé
— — — Portillo
— — — Randy Newman
— — — Robert Cray Band
— — — Robert de Niro
— — — Rony Seikaly
— — — Ry Cooder
— — — Steve Smith
— — — The Band
— — — Willie Nelson
— — letras
— — — Curier
— — — letras residentes
— — — — Currier
— — — — Orator
— — — — Roman
— — — — Sans Serif
— — — — Scritp
— — — Roman
— — — Sans Serif
— — — Serif
— operações mentais
— polifonia poética
— pulsão
— sublimação
— times de qualidade
— transformações
— verso harmônico
— verso melódico

ANEXO 10: Taxonomia de *instituições* (corpus genérico)

INSTITUIÇÕES
— Banco da Amazônia
— Banco de o Brasil
— Banco do Brasil
— Banco do Nordeste
— banco Meridional
— Caixa Econômica Federal
— empresas
— — Água de Cheiro
— — Alcoa
— — Alpargatas
— — AM / PM
— — Andrade Gutierrez
— — Arbi
— — AT & T
— — Banco Francês e Brasileiro
— — Banco Nacional
— — Banco Noroeste
— — Banco Real
— — Boeing
— — Boston de o Brasil
— — Brasif Comercial
— — Brittish Petroleum
— — Caesar Park Hotel
— — Carrefour
— — Chrysler
— — Citibank
— — Citrovita
— — Coca-Cola
— — Coelho
— — Compton's Nem Media
— — Discis Knowledge Research
— — Docol
— — Dupont
— — Flytour
— — Ford
— — Glaxo
— — grupo Gerdau
— — instituições de pesquisa
— — Interpass Club
— — Itambé
— — Jacadi
— — Kurzweil Music Systems
— — Lloyds Bank
— — Moinho Santista
— — Montreal Informática
— — Nacional Seguros
— — Nestlé
— — Norrau Informática
— — Pantanal
— — Papel Simão
— — Parmalat
— — Pinguim
— — Pirelli
— — Rio-Sul
— — Rummler-Brache Group
— — Sanbra
— — Santa Celina Mineradora
— — Shell
— — Souza Cruz
— — Stella Barros Turismo
— — TAM
— — Telerj
— — Tintas Coral
— — universidades
— — Varig
— — Vicunha
— — Xerox
— escolas
— — escolas de samba
— — — Mangueira
— — — Portela
— — — Salgueiro
— instituições públicas
— — cinemas
— — salas de convenções
— — teatros
— universidades
— — empresas excelentes
— — — RBS
— — — Varga
— — — WEG

ANEXO 11: Taxonomia de *jogadores* (corpus genérico)

JOGADORES
— Alexi Lalas
— Almir
— Asprilla
— Axel
— Bebeto
— Boiadeiro
— Branco
— Cafu
— Careca
— César Sampaio
— Cobi Jones
— Cuca
— Dener
— Edmundo
— Escobar
— Gary Lineker
— Gilmar
— Guga
— Jairzinho
— Jim Courier
— Toninho Cerezo
— Valderrama
— Valência
— Velloso
— Vernon Maxwell
— Viola
— Zenon
— Zico
— Zinho
— jogadores reservas
— — Edna
— — Fofão
— — Popó
— — Virna
— jogadores de nomes complicados
— — Beschastnykh
— — Jin Ho Cho
— — Mbouh Mbouh
— jogadores em Portugal
— — Aldair
— — Mozer
— jogadores profissionais
— — Ézio
— — Leonardo
— — Lira
— — Marquinhos
— — Renato Gaúcho
— — Torres
— — Válber
— — Valdir
— — Viola
— Kenny Smith
— Leonardo
— Luisinho
— Mazinho
— Michael Stich
— Moeller
— Mozer
— Muller
— Otis Thorpe
— Paulo Roberto
— Perea
— Pete Sampras
— Pierre Littbarski
— Prosinecki
— Raí
— Renato
— Ricardo Gomes
— Ricardo Rocha
— Riedle
— Rincón
— Rivaldo
— Roberto Dinamite
— Romário
— Ronaldo
— — titular
— Souza
— Tab Ramos

ANEXO 12: Relações cujo hipônimo é um nome próprio (corpus saúde)

LDL<proteína	A República<diálogos
Leplat<autores	AACD<instituições
Londres<cidades populosas	AAS<analgésicos a base de ácido acetil salicílico
Londrina<municípios	Abastecimento de Água<indicadores
LSD<drogas	ADC<métodos
LSD25<alucinógenos	Aedes albopicutus<espécies
Lutzomyia longipalpis<mosquitos flebótomos	AIDS<condições crônicas
Malásia<países	Aids<doenças
Martins<autores	Alzheimer<doenças
Marvin Harris<autores	América<regiões do planeta
matrícula SIAPE<identificação	Apae<instituições tradicionais
Medellín<cidades violentas	Aspergillus<fungos
Mengele<alemães	Associação dos Extratores<organizações locais
México<países	Baggio<estudiosos
Michel de Montaigne<pensadores	Bambuí<municípios de pequeno porte
Minas Gerais<estados	Bangladesh<nações
Ministério da Saúde<convidados	Barash & Weinstein<autores
Mogi das Cruzes<municípios da área metropolitana	Baruch Spinosa<filósofos
	Bom Jesus<bairros
Morbillivirus<vírus	Brasília<metrópoles
MSX 1 UM<gene	British American Tobacco<companhias
Município de Campinas<municípios vizinhos	C.glabrata<espécies de Candida
Mycobacterium tuberculosis<bactéria	Cali<cidades violentas
Nestlé<laboratórios multinacionais	Cármides<diálogos
NIRH<programas	Casa Vital Brazil<fundação
Olinda<cidades de o interior pernambucano	Cássia dos Coqueiros<municípios
OMS<organizações internacionais	Cazaquistão<países
ONGs<atores sociais	Ceará<estados
Optalidon<medicamentos	Cedau<programas
Panstrongylus megistus<espécies	Centro de Atenção Crônica<organizações não governamentais
Papaver somniferum<planta	
Partenon<bairros	CMV<característica de agentes virais
Peak Flow Meter<aparelho	Conselho Tutelar<instituições públicas
Penicillium chrysogenum<fungos	Criptococcus<fungos
	Philip Morris<companhias

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo