

## Subsídios para a Elaboração Automática de Taxonomias

Maria Cláudia de Freitas<sup>1</sup>, Violeta Quental<sup>1</sup>

<sup>1</sup>Departamento de Letras – Pontifícia Universidade Católica do Rio de Janeiro (PUC/RJ)  
Rio de Janeiro – RJ – Brasil

{claudiaf,violetaq}@let.puc-rio.br

**Abstract.** *This paper presents linguistic resources to the automatic building of corpus-based taxonomies. These resources are applied to a domain corpus and to a generic corpus. The comparison between them shows that producing inferences in order to construct taxonomies can be of great value to build domain specific taxonomies.*

**Resumo.** *Neste trabalho, apresentamos subsídios lingüísticos para a construção automática de taxonomias a partir de corpus. Tais subsídios são aplicados em um corpus específico quanto ao domínio e em um corpus genérico. A comparação revela que, embora o cruzamento de relações de hiperonímia, capaz de produzir inferências e construir taxonomias, não seja uma técnica explorada freqüentemente, seus resultados são de grande valia para a elaboração de taxonomias de domínio.*

### 1. Introdução

Em geral, trabalhos que envolvem a extração automática de relações de hiponímia a partir de corpus (Hearst 1992, 1998; Cederberg e Widdows, 2003) não utilizam os resultados dessa extração para a realização de inferências. Isto é, se são extraídas as relações "abacaxi é uma fruta" e "fruta é um alimento", não há cruzamento das informações para que se produza, automaticamente, "abacaxi é um alimento". Uma possível explicação para esse descarte é a grande quantidade de erros gerada, principalmente se as relações são extraídas de corpora gerais quanto ao domínio, como é o caso de corpora compostos por textos jornalísticos. Kilgarriff (2003) se opõe à utilização de tesouros 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, e não em palavras. Nesse sentido, inferências seriam 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*. De fato, este é um passo delicado, uma vez que inferências pressupõem um significado fixo e estável das palavras. Porém, em favor de uma taxonomia baseada em palavras, argumentamos 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, invocamos a restrição "one sense per discourse" (Yarowsky 1995), segundo a qual o significado de uma dada palavra é altamente consistente em um determinado discurso.

Neste trabalho, propomos a produção de inferências, e conseqüente elaboração de taxonomias, a partir da extração de relações de hiperonímia do corpus. A extração toma por base alguns dos padrões léxico-sintáticos apresentados em Hearst (1992, 1998), adaptados para o português, além de dois outros padrões por nós encontrados no corpus. Em seguida, os resultados da extração são cruzados, tendo em vista a construção da taxonomia. Por fim, os resultados obtidos a partir de um corpus específico de domínio são comparados com os obtidos a partir de um corpus genérico, majoritariamente jornalístico, com o objetivo de avaliar a eficácia do processo.

O restante do artigo está organizado da seguinte maneira: na seção 2 são brevemente apresentados alguns trabalhos que tratam da extração automática de hiperonímia. A seção 3 descreve a metodologia utilizada e a seção 4 contém nossas considerações finais.

## 2. Trabalhos Relacionados

Marti Hearst (1992, 1998) foi a primeira a aplicar a idéia de que determinados padrões léxico-sintáticos poderiam, sistematicamente, expressar determinadas relações semânticas no PLN. Especificamente, Hearst (1992, 1998)<sup>1</sup> 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. Como a maioria dos termos da WordNet são substantivos sem (ou com um único) modificadores, os algoritmos de Hearst objetivam extrair apenas relações entre nomes sem modificadores, tanto no sintagma hiperônimo quanto no hipônimo. No trabalho de 1998, 200 instâncias do padrão "e outros" foram avaliadas manualmente. 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.

Morin e Jacquemin (2004) apresentam um sistema – Prométhée – que extrai e utiliza padrões léxico-sintáticos no estilo Hearst a partir de corpus. Esses padrões foram aplicados em um corpus constituído de resumos e títulos de artigos científicos, o corpus "[AGRO-ALIM]". A avaliação dos pares extraídos mostrou uma alta qualidade das relações produzidas, com uma precisão de 82% e uma abrangência de 56%.

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 e, ao final, conjugando os padrões de Hearst com técnicas estatísticas, atingem 64% de acertos, tomando por base a avaliação manual de 260 relações.

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. Por outro lado, a metodologia apresenta a grande

---

<sup>1</sup> A principal diferença entre os dois trabalhos está no corpus utilizado: em 1992, os padrões foram extraídos da *Grolier's Encyclopaedia*; em 1998, de seis meses do jornal *New York Times*.

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 os resultados de Morin e Jacquemin (2004) são, em termos gerais, bastante superiores aos de Hearst e de Cederberg e Widdows. Porém, uma comparação exata entre os trabalhos não é possível por uma série de 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 há informações sobre 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.

### 3. Metodologia

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 coletados na Internet. Para a aplicação dos algoritmos de identificação de padrões sobre o corpus, é necessário um corpus com etiquetas de (i) classes de palavras e (ii) sintagmas nominais. Para atender à primeira exigência, o corpus foi processado pelo parser PALAVRAS (Bick, 2000), na opção “*morphological tagging*”. Em seguida, para a etiquetagem de sintagmas nominais, o corpus passou pelo identificador de SNs descrito em (Santos e Oliveira, 2005)<sup>2</sup>.

Após o processo de etiquetagem automática, o corpus foi manualmente revisto. Em seguida, regras baseadas em expressões regulares foram aplicadas ao corpus, a fim de extrair as relações de hiperonímia. As regras estão descritas nas próximas seções.

---

<sup>2</sup> Como o identificador de SNs de Santos e Oliveira (2005) foi treinado com as etiquetas gramaticais do Lácio-Web, que são diferentes das etiquetas gramaticais do Palavras, foi ainda necessária uma etapa intermediária de conversão de etiquetas Palavras – Lácio.

### 3.1. Descrição dos padrões

#### 3.1.1. 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, como ilustram (1) e (2). 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 com isso 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.

(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,....

Além disso, o “como” que nos interessa quase não aparece nas gramáticas: trata-se de um “como” que pode ser classificado como uma “palavra denotativa”, do mesmo modo que seria a expressão “por exemplo”. Ou 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<sup>3</sup>, 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.

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 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 reconhecê-lo.

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 (3) são extraídas as relações (3a) e (3b).

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

(3a) Salmonella < infecções por bactérias

(3b) Shighella < infecções por bactérias

A solução foi criar, além do SN hiperônimo (SN Hiper), o SN HHiper, que considera como SN hiperônimo o primeiro N à esquerda do “como/tais como”. A análise do corpus mostrou, porém, que uma outra alteração na regra permitiria ainda

<sup>3</sup> No Palavras (Bick, 2000), o “como” da frase (2) recebe a etiqueta ADV @AS-N<, que é interpretada como uma construção elíptica – uma oração adverbial em que o verbo ser está elíptico: “...outros mamíferos *como* [o são] camundongos, bezerros....” Porém, esta etiqueta só aparece quando é realizado o parsing morfosintático completo. Quando se escolhe a opção “*morphological tagging*” (como neste trabalho), a etiqueta recebida é <rel> <ks> <prp> ADV – a mesma etiqueta que os demais “como”\_advérbios recebem.

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". A regra final utilizada na identificação do padrão "como/tais como" foi, portanto, desmembrada em duas, em que HHiper representa o padrão em que se considera hiperônimo o primeiro substantivo à esquerda e Hiper representa o padrão em que todo o SN será extraído como hiperônimo:

(Ia) SN HHiper (tais como | como\_PDEN)  $SN_1 \{ , SN_2 \dots , \}$  (e | ou)  $SN_i$

(Ib) SN Hiper , (tais como | como\_PDEN)  $SN_1 \{ , SN_2 \dots , \}$  (e | ou)  $SN_i$

### 3.1.2. O padrão “e outros”

A identificação das relações expressas pelo padrão "e outros", tratado em Hearst (1998), também sofre com problemas decorrentes da ambigüidade do sintagma preposicionado, como ilustra (4):

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

Neste caso, porém, a dificuldade de segmentação não está no SN hiperônimo, mas nos SNs hipônimos. A solução que encontramos para minimizar o problema foi criar, ao lado do SN HHiper, o SN HHipo: é considerado SN hipônimo o primeiro substantivo 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. Assim, para o padrão "e outros", a regra utilizada foi

(II) SN HHipo  $\{ , SN_{Hipo_i} \} * \{ , \}$  elou outros SN Hiper.

### 3.1.3. O padrão “tipos de”

A partir da observação do corpus, percebemos que o padrão "tipos de" também expressa relação de hiperonímia. A regra correspondente ao padrão é:

(III) *tipos de* SN Hiper:  $SN_1 \{ , SN_2 \dots , \}$  (e | ou)  $SN_i$

### 3.1.4. O padrão “chamado”

Este padrão também foi descoberto a partir da observação do corpus. Nele, também há dificuldade na identificação da relação em decorrência da ambigüidade do sintagma preposicionado, como ilustra (5) e, novamente, foi utilizada a estrutura HHiper. A regra final para a identificação está descrita em (IV), em que só o último SN é considerado.

(5) ... [a alta frequência da doença mental] chamada [esquizofrenia].

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

## 3.2. Resultados da Extração

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, partimos do pressuposto de que os padrões investigados expressam, de fato, relações de hiperonímia, ainda que não constituam relações "convencionais" de um ponto de vista lexicográfico. Desse modo, consideramos corretas relações como

sensibilidade<condição

reforma de um jardim<trabalhos voluntários

Consideramos erros apenas casos em que a relação extraída não estava correta devido (i) à ambigüidade do sintagma preposicionado; (ii) à presença de uma estrutura adverbial deslocada da ordem direta ou encaixada; (iii) à elipse de algum termo ou (iv) à presença de uma oração no interior do sintagma hiperônimo ou hipônimo. Com esses critérios, a análise manual das relações extraídas revelou um índice de 76.4% acertos. Porém, embora coerente com o ponto de vista teórico de inspiração wittgensteiniana assumido (Wittgenstein 1953), 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 num dado contexto de uso, é pouco significativa na elaboração de uma taxonomia e pode ser eliminada sem prejuízo (ou com prejuízo mínimo) de informação.

### 3.3. Validação

A segunda etapa da avaliação teve como objetivo tornar os resultados "mais comparáveis" e "mais significativos": avaliadores<sup>4</sup> fizeram a validação de uma amostra dos resultados considerados "corretos". Das 2244 relações corretamente extraídas – assumindo o critério puramente sintático –, uma amostra aleatória de 436 relações (cerca de 1/3) foi selecionada para avaliação. 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, frequentemente 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 "ser útil". Os resultados da avaliação humana estão na tabela 1.

---

<sup>4</sup> Participaram desta etapa três 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.



**Tabela 1: Resultados da avaliação humana**

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

Os resultados da avaliação indicam que a maior parte dos erros está na categoria 1, sendo decorrência de definições gerais demais ou específicas demais – e, conseqüentemente, pouco úteis. É o caso de relações cujo hiperônimo é um substantivo do tipo “fator”, “termo” “elemento”, “questão”, “aspecto”, entre outros. Tais hiperônimos se enquadram na lista de substantivos genéricos descritos em Marques (1995), e de substantivos-suporte descritos em Oliveira (2006). De modo a eliminar tais relações gerais demais e pouco informativas, aplicamos um filtro para eliminar as relações cujo hiperônimo fôsse um substantivo genérico/suporte. Para diminuirmos os erros da categoria 2, relativos principalmente à “dependência contextual” de algumas relações, aplicamos mais dois outros filtros: um para eliminação de pronomes dêiticos e outro para eliminação de alguns adjetivos que Hearst chama de adjetivos comparativos, como “importante” e “menor”. Após a aplicação dos filtros, o número de relações extraídas caiu de 2241 para 1937. Dessas, uma amostra aleatória de 430 foram avaliadas manualmente. Os novos resultados estão na tabela 2.

**Tabela 2: Resultados da avaliação humana após a aplicação de filtros**

<b>Classificação</b>	<b>Qtd de relações COM filtro</b>	<b>Qtd de relações SEM filtro</b>
<b>3</b>	<b>349 (81%)</b>	320 (73.4%)
<b>2</b>	<b>28 (6.5%)</b>	15 (3.4%)
<b>1</b>	<b>20 (4.6%)</b>	70 (16%)
<b>0</b>	<b>33 (7.6%)</b>	31 (7.1%)

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). Uma nova amostra aleatória de 527 relações foi analisada manualmente, e os resultados estão na tabela 3.

Embora o índice de acertos (75%) seja inferior ao obtido 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).

**Tabela 3: Resultados com o corpus genérico**

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

### 3.4. Produção de Inferências

Com o cruzamento das informações obtidas na extração dos padrões léxico-sintáticos no corpus de saúde, foram encontradas 420 taxonomias. Uma amostra aleatória de 140 taxonomias foi avaliada manualmente. Surpreendentemente, encontramos erros em apenas 14 taxonomias, o que significa um total de 90% de acertos, e contradiz a posição de Kilgarriff (2003). Por outro lado, esse alto índice de acertos se deve, em grande parte, à utilização de um domínio restrito e técnico, que dá pouca margem à ocorrência de variações entre os significados, como ilustra a taxonomia de “artrópode”, abaixo:

Artrópodes  
 --ácaros--  
 --carrapatos  
 --piolhos  
 --pulgas  
 --mosquitos  
 ----mosquitos flebótomos  
 ----Lutzomyia longipalpis

A fim de verificar se o alto índice de acertos obtido na realização de inferências foi consequência 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 142.000 palavras. Foram produzidas 920 taxonomias. Dessas, uma amostra aleatória de 50 foi avaliada manualmente. O índice de acertos caiu para 60%. 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. O exemplo abaixo, da taxonomia de “adornos”, ilustra o fato.

Adornos  
 --anjos  
 --cavalos-marinheiros  
 --estrelas  
 ---Aretha Franklin  
 ---B.B. King  
 ---Catherine Deneuve



### 3.5. Discussão dos Resultados

Na comparação entre as taxonomias produzidas, uma primeira observação diz respeito ao alto número de taxonomias geradas a partir do corpus genérico (920), principalmente se considerarmos que o corpus de saúde, com quase 2 milhões de palavras, produziu “apenas” 420 taxonomias. Essa proliferação excessiva de taxonomias no corpus genérico é consequência de dois fatores: (i) o caráter geral do corpus CETENFolha, que trata de uma vasta gama de assuntos; (ii) 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 taxonomia de domínio e a taxonomia genérica 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 categorias “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.

### 4. Considerações Finais

Apresentamos aqui subsídios para a elaboração automática de taxonomias. 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 em Hearst (1992), acreditamos 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.

### 5. Referências

- AIRES, R.V.X.; ALUÍSIO, S.M. Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente. Relatórios do NILC, NILC-TR-01-8, 2001.
- BICK, E. The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD Thesis. Aarhus University, 2000.

- 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, Canadá, 111-118, 2003.
- CRUSE, D. Lexical Semantics. Cambridge, Inglaterra: Cambridge University Press, 1986.
- HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, 1992.
- HEARST, M. Automated discovery of WordNet relations. In: Fellbaum, Christiane, ed., WordNet: An Electronic Lexical Database. MIT Press, 1998.
- KILGARRIFF, A. Thesauruses for Natural Language Processing. In: Proceedings of NLPKE, Beijing, China, p.5-1, 2003.
- LYONS, J. Semântica. Martins Fontes, 1980.
- MARQUES, M. H. D.. Léxico de alta frequência na língua portuguesa. In: HEYE, J. (org). Flores verbais, uma homenagem lingüística e literária para Eneida do Rego Monteiro Bomfim no seu 70º aniversário. Rio de Janeiro: 34 Editora, p. 247-282, 1995.
- MORIN, E. e JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. In: Computer and the Humanities, vol. 38 (4), 343-362, 2004.
- 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.
- SANTOS, C.N., OLIVEIRA, C.M. Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: Anais do XXV Congresso da Sociedade Brasileira de Computação, Brasil, 2005.
- SNOW, R., JURAFSKY, D., e NG, A. Y. Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17, 2005.
- YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, pp 189-196, 1995. Boulic, R. and Renault, O. (1991) "3D Hierarchies for Animation", In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons Ltd., England.
- WITTGENSTEIN, L. Investigações Filosóficas. Coleção Os Pensadores, São Paulo: Abril Cultural, 1979.