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* .

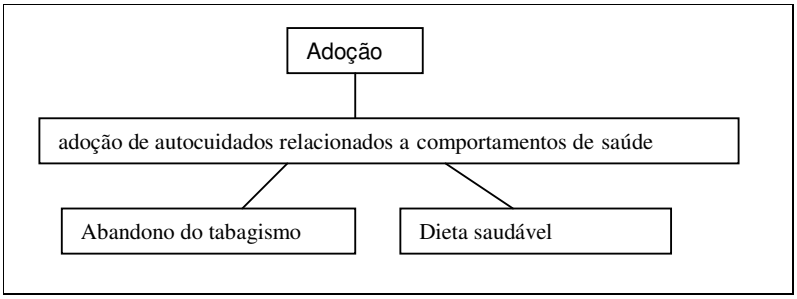

Figura 3: Taxonomia de adoção produzida pela regra "hiperN"

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*.

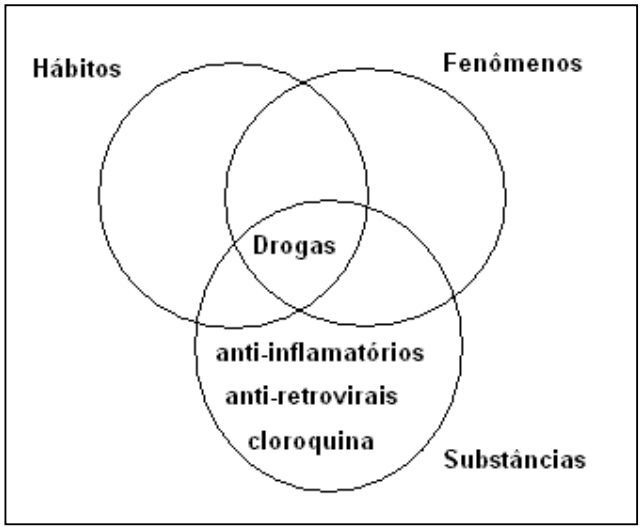

Figura 7: Diferentes contextos de uso de *drogas*

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.

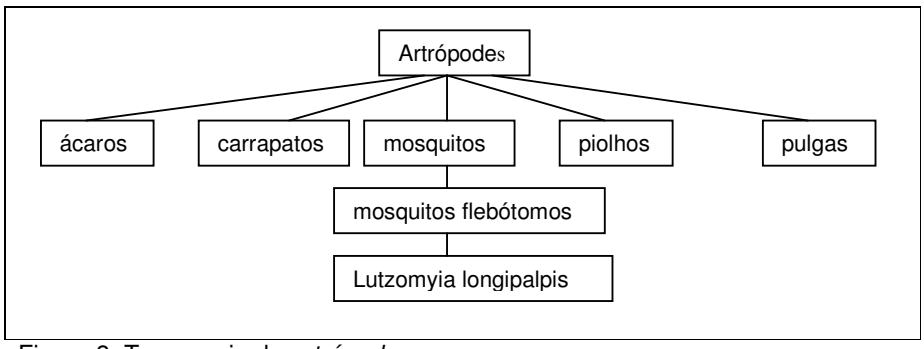

Figura 8: Taxonomia de *artrópodes*

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.

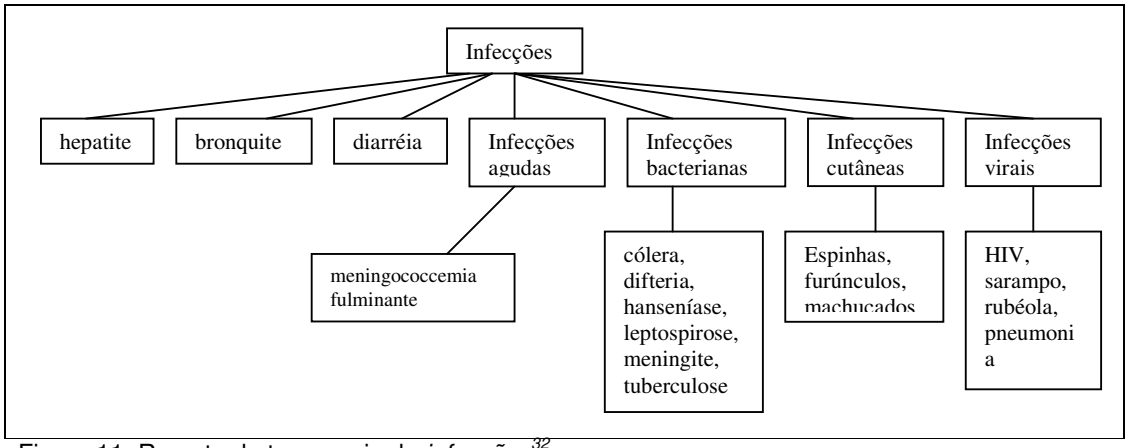

Figura 11: Recorte da taxonomia de *infecções*[32]

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).

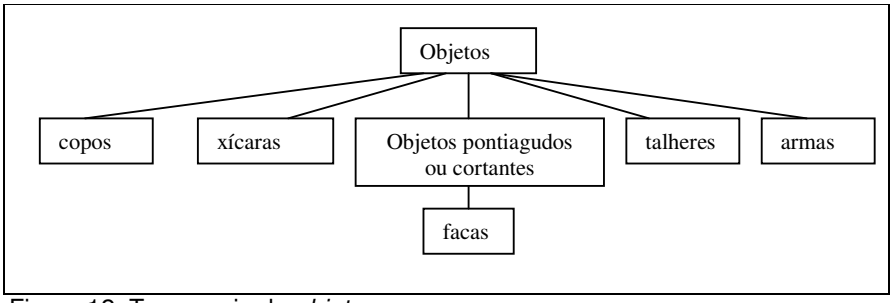

Figura 12: Taxonomia de *objetos*

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).

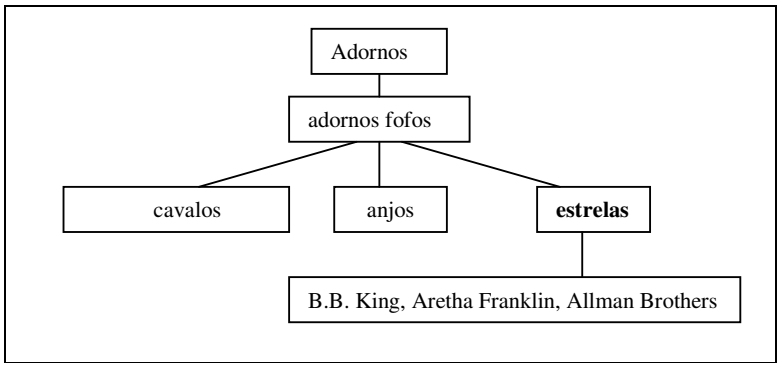

Figura 13: Taxonomia de *adornos*

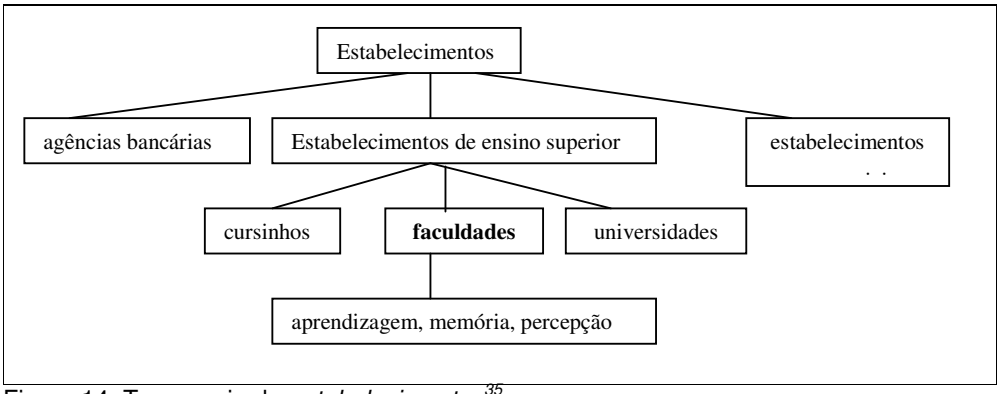

Figura 14: Taxonomia de *estabelecimentos*[35]

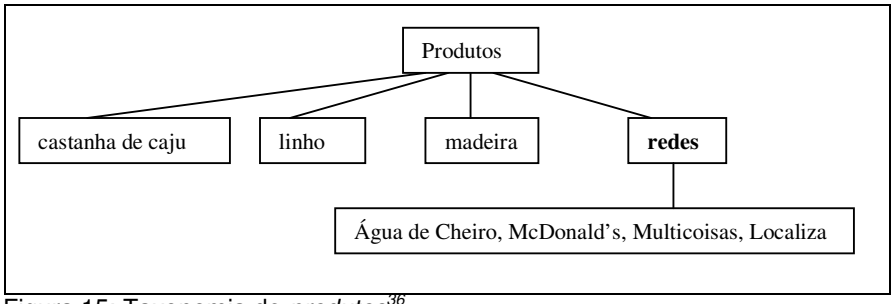

Figura 15: Taxonomia de *produtos*[36]

[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*