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,

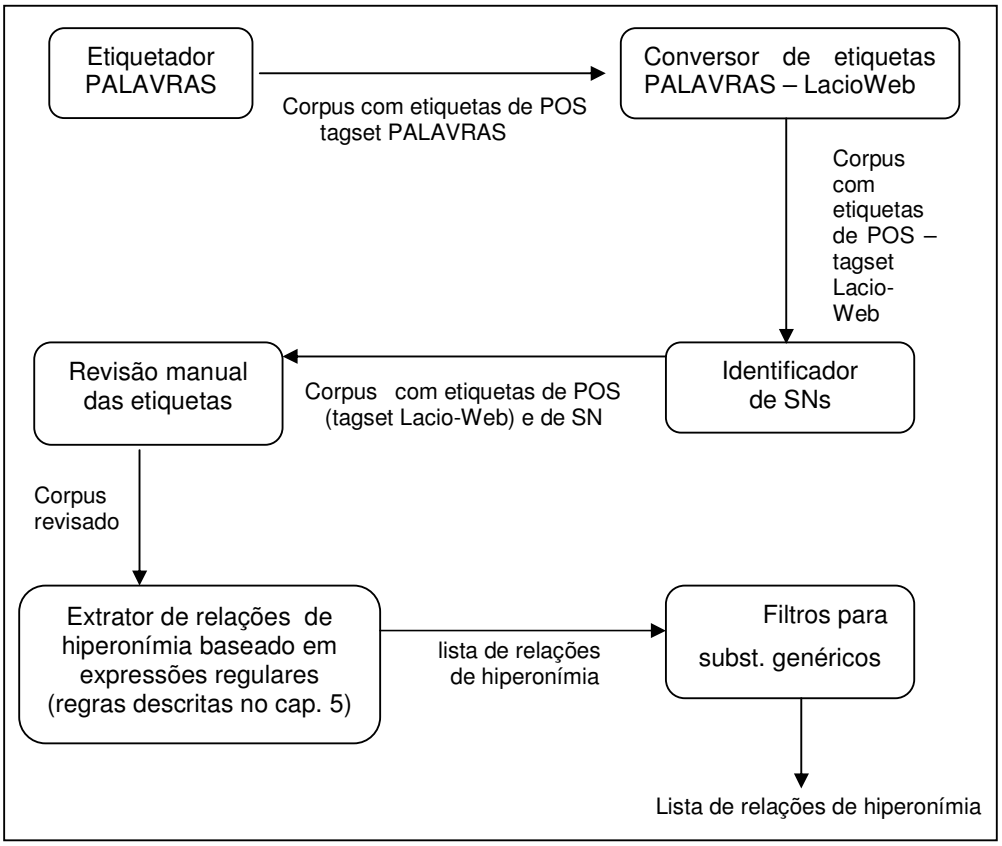
Entre os idosos examinados, houve alguns participantes que, por problemas clínicos, ***tais como*** *visão, audição, entre outros, não conseguiram completar...*

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.