5
Metodologia

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

5.1.
O corpus

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

5.1.1.
O pré-processamento do corpus

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

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

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

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

5.2. Descrição dos padrões

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

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

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

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

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

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

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

As seções seguintes detalham cada um dos padrões utilizados.

5.2.1. O padrão "tais como"

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

Deste modo, para o padrão original

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

utilizamos, para o português,

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

capaz de extrair relações de estruturas como

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.2. O padrão "e/ou outros"

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.3. O padrão "tipos de"

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

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

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

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

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

5.2.4.
O padrão "chamado/a/os/as"

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

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

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

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

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

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

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

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

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

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

(29') angina = doença do beijo

(30') tétano neonataltétano = mal dos sete dias

Para a identificação automática desta estrutura, a regra utilizada foi

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

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

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

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

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

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

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

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

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

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

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