Os padrões de Hearst como recursos auxiliares semiautomáticos para a eficácia na leitura documentária

Walter Moreira
José Carlos Francisco dos Santos
Érica Fernanda Vitorini

Como citar: MOREIRA, W.; SANTOS, J. C. F. dos; VITORINI, E. F. Os padrões de Hearst como recursos auxiliares semiautomáticos para a eficácia na leitura documentária. *In*: FUJITA, M. S. L.; NEVES, D. A. de B.; DAL'EVEDOVE, P. R. (org.). **Leitura documentária**: estudos avançados para a indexação. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2017. p. 157-194.
DOI: https://doi.org/10.36311/2017.978-85-7983-917-7.p157-194

Os padrões de Hearst como recursos auxiliares semiautomáticos para a eficácia na leitura documentária

Walter Moreira
José Carlos Francisco dos Santos
Érica Fernanda Vitorini

1 Introdução

Como ponto de partida para a discussão que se faz neste capítulo, considera-se como imperativo a aplicação de metodologias de tratamento temático no amplo e variado conjunto da produção científica. A confluência das variáveis tempo destinado à análise dos documentos em sistemas de informação documentária e quantidade/qualidade destes documentos requer que o processo de tratamento temático seja realizado de modo eficaz, com o aproveitamento máximo dos recursos aplicados.

O tratamento temático da informação e, mais especificamente, a análise documentária é, por sua própria natureza intelectual, dispendioso. Desse modo, é fundamental que a análise documentária seja realizada com a observação atenta de princípios metodológicos e com o recurso de ferramentas que possam torná-la mais eficaz e racional. Entretanto, como adverte Kobashi (1994, p. 22), a análise documentária

https://doi.org/10.36311/2017.978-85-7983-917-7.p157-194

> não se reduz a um conjunto de regras perenes, utilizáveis em todas as circunstâncias. Ela é, antes de tudo, uma disciplina de natureza metodológica que, para avançarteórica e praticamente, deve criticar continuamente seus pressupostos, procedimentos e instrumentos; deve, ao mesmo tempo, com base na reflexão permanente, elaborar novas hipóteses de trabalho que contribuam para aperfeiçoar os processos que lhe dizem respeito.

A análise documentária diz respeito ao exame do documento visando a sua representação para fins de recuperação da informação. Desse modo, a análise documentária configura-se como metodologia por meio da qual o conteúdo de um determinado texto é interpretado, condensado e lhe são atribuídos descritores, tendo em vista sua representação para fins de indexação.

A representação é um produto diferente do texto original, mas que guarda com ele relações de equivalência e de contiguidade. A construção da representação documentária é marcada por um processo de redução crescente, assim caminha-se

> do texto para o resumo, do resumo para o enunciado, do enunciado para a unidade de tradução via códigodocumentário. A atividade de AD caracteriza-se, portanto, como uma sucessão de processos de transformação do texto original, observando-se, a cada etapa, graus crescentes de generalização" (LARA, 1993, p. 41).

Nos modelos mais tradicionais dos sistemas documentários, esta representação documentária ocorre na forma de classificação, de indexação ou de resumo e tais representações visam a qualificar o processo de recuperação da informação pelo usuário.

Conforme Lara (2009, p. 28), o termo "análise documentária" foi criado por Jean-Claude Gardin para "designar as operaçõessemânticas que transformam um texto original em uma ou várias palavras-chave, ou ainda, paráfrases, visando facilitar a representação [...] de 'conteúdos' e a recuperação da informação".

No Brasil, as pesquisas desenvolvidas pelo Grupo Temma, que passa a existir formalmente com a publicação do livro "Análise documen-

tária: a análise da síntese", em 1987 (Lara, 2009), formam grande parte do estofo teórico e metodológico das pesquisas sobre análise documentária.

Para que se efetue a análise documentária, requer-se a realização de leitura documentária, uma modalidade de leitura profissional que consiste na análise do conteúdo do documento visando a identificação e a distinção das informações essenciais e acessórias, sendo que tal distinção ocorre conforme os interesses relativos ao contexto em que se produz a análise. Utilizando-se de técnicas de condensação documentária adequadas à modalidade de representação que se deseja construir, as informações essenciais são reelaboradas como representações documentárias e passam a compor, nessa condição, sistemas de informação documentária.

A leitura documentária, como qualquer atividade de leitura, é um processo de produção de sentido que tem como ponto de partida o texto. Neste caso há, contudo, uma variável que não está presente na leitura que se realiza para fins de lazer, por exemplo: o leitor profissional, o que realiza a leitura documentária, não é, normalmente, previsto como leitor pelo autor. Nas palavras de Lara (2009, p. 34): "frente ao documento, o leitor-documentalista não se caracteriza como um leitor-modelo, uma vez que não dispõe, necessariamente, de condições para estabelecer com o texto uma negociação". Além do mais, acrescenta-se o fator urgência (ou tempo reduzido), que obriga à realização de leitura mais rápida e seletiva do se poderia esperar a partir de um modelo convencional. Ainda citando Lara (2009, p. 34): "o leitor-documentalista realiza uma leitura que se enquadra num processo de produção industrial de textos (parafraseando Gardin), não podendo dedicar mais tempo à leitura do que aquele previsto na atividade de indexação de um grande volume de publicações".

Em busca de uma metodologia para a leitura documentária, Gil Leiva e Fujita (2012), estabelecem alguns procedimentos orientadores. Rubi, Fujita e Boccato (2012, p. 224) também caracterizam a leitura documentária e apresentam uma metodologia para a sua realização que envolve a análise mais cuidadosa dos seguintes elementos: a) introdução: com foco nos objetivos; b) "leitura das frases introdutórias de parágrafos e capítulos"; c) procedimentos metodológicos, contemplando "técnicas, instrumentos, procedimentos adotados na realização da pesquisa, bem como o e ambiência em que esta se passa"; d) conclusão em sua relação

com os objetivos propostos; e) "gráficos, tabelas, diferenciação tipográfica etc."; f) título, subtítulo, resumo e palavras-chave. Estes elementos devem ser verificados ao final da leitura, segundo as autoras, para que não a direcione de algum modo.

A leitura documentária é uma modalidade de leitura profissional, mais especificamente é a leitura realizada pelo indexador em contexto profissional. Trata-se de uma modalidade de leitura que visa a um fim específico que é construir a representação documentária. Não soa correto, portanto, do ponto de vista da organização lógica dos conceitos, tomar os termos "leitura documentária" e "leitura profissional" como sinônimos, em relação de equivalência. Desse modo, a expressão "leitura profissional", quando se refere à atividade realizada pelo indexador, deve aparecer adjetivada, como ocorre em Silveira e Moura (2007), com a expressão "bibliotecário-indexador" ou em Redígolo e Fujita (2015), que utilizam a expressão "leitura profissional do catalogador", cuja leitura profissional viabiliza a "consecução de seus objetivos de síntese e seleção de conceitos" (REDÍGOLO; FUJITA, 2015, p. 367)

Observado como leitor "comum", o indexador utiliza-se de recursos também "comuns" a qualquer outro leitor, acionando

> o processamento humano de informações, realizado com a memória de curto prazo (*input* visual), a memória de longo prazo (esquemas e conhecimento prévio) e as habilidades operatórias de pensamento (análise e síntese). [Observado como] leitor profissional é considerado a partir da perspectiva de seu contexto, atuação e formação profissional" (FUJITA; RUBI, 2006, p. 1).

Deste modo, são as técnicas empregadas na leitura que vão diferenciar o leitor "comum" do leitor profissional. De outro lado, é a tipologia de recursos empregados que irá diferenciar a leitura documentária no conjunto das leituras profissionais.

No trabalho em que sistematiza e apresenta uma abordagem teórico-metodológica para a construção de informações documentárias (especialmente o resumo), Kobashi (1994), quando discute a relação entre a prática documentária e o seu processamento automatizado, aponta que, naquele contexto, sem o conhecimento profundo das operações relativas

ao tratamento da informação documentária, os recursos da computação eram subutilizados. Citando-a textualmente, pode-se ler:

> Os fracassos [da incorporação da informática às tarefas da biblioteconomia/documentação] não se deveram unicamente à limitação das máquinas. As experiências revelaram que, muitas vezes, as dificuldades provinham de lacunas da própriaárea: especificamente, da ausência de conhecimentos suficientemente sistematizados sobre operações por vezes básicas e elementares do cotidiano profissional. Apoiada fortemente no conhecimento empírico, na intuição e no habito, a área, por um longo período, rejeitou a teoria, considerando-a supérflua.

O quadro não é mais o mesmo e houve evolução, naturalmente. Alguns autores, dentre eles Kobashi, no mesmo trabalho citado (1994) e Gil Leiva e Fujita (2012), forneceram importantes subsídios teóricos e metodológicos relativos aos processos inerentes à análise documentária. Sobre este aspecto, apontam-se apenas alguns outros trabalhos seminais: Smit (1989), Guimarães (1994), Lara (1999), Lucas (2000), Pinto Molina (2001), Fujita (2003), Montesi (2006) e Dias e Naves (2013).

A análise documentária, visa, como se disse anteriormente, a construção de representações documentárias que possam ser utilizadas tanto pelo indexador quanto pelo usuário pesquisador para fins de diálogos com o sistema. Neste caso, como se trata de uma linguagem construída para fins de comunicação entre usuários (leia-se indexador e pesquisador) e sistema, ou seja, uma linguagem documentária, é preciso que seja aplicado ao conjunto das representações documentárias metodologias de controle de vocabulário.

Para Lancaster (2002) o controle do vocabulário tem os objetivos de facilitar a representação dos assuntos tanto para os profissionais como para os usuários; agrupando os sinônimos e os quase sinônimos, diferenciando os homógrafos e relacionando os termos com significados próximos, sendo assim tem como fundamento diminuir essas diferenças e padronizar o seu uso facilitando, consequentemente, o acesso.

Deste modo, acredita-se que uma representação de qualidade pode proporcionar uma recuperação eficaz da informação. Problemas no processo de representação, como a ausência de um vocabulário controlado,

sua falta de atualização ou até mesmo um vocabulário que não represente a linguagem da comunidade usuária, geram falhas nessa comunicação e concorrem para a insatisfação na recuperação da informação. Além de comprometer a qualidade das buscas e do processo de recuperação da informação como um todo, a incompatibilidade entre a linguagem documentária e a linguagem do usuário, adverte Boccato (2009, p 21), compromete "a atuação do bibliotecário na representação dos conteúdos documentários no processo de indexação, o usuário na realização das buscas satisfatórias desses conteúdos no processo de recuperação da informação e, consequentemente, a credibilidade dos sistemas".

Embora se trate de um tema que não será explorado neste capítulo, é preciso reforçar o argumento de que não será possível estabelecer um diálogo eficaz entre usuários, linguagens documentárias, sistemas de informação documentária e indexadores (bibliotecários) sem a adoção de políticas de indexação coerentes e consistentes. A este respeito, pode-se consultar o já clássico artigo de Carneiro (1985) e os trabalhos de Rubi (2004; 2008), Gil Leiva e Fujita (2012) e Fujita (2016).

A política de indexação funciona como eixo orientador para o profissional indexador no momento da análise documentária, pois por meio dela é possível estabelecer o perfil da instituição e da comunidade usuária, entre outros critérios, e assim definir os parâmetros orientadores para a leitura e seleção dos descritores mais adequados à representação documentária.

A construção eficaz da representação como resultado da análise documentária requer, portanto, como condição *sine qua non*, um leitor orientado por uma política de indexação que: a) seja consciente a respeito da complexidade do ato de ler e que esteja munido de estratégias metacognitivas de leitura (um leitor profissional proficiente); b) conheça o perfil de interesse da comunidade usuária a que atende e c) conheça os diversos modelos de organização textual e saiba aplicar recursos adequados à identificação e seleção de informação essencial observando-se as diferentes tipologias de estrutura textual.

2 Organização da informação textual

O ponto de partida para a realização da análise documentária, conforme a compreensão do Grupo Temma relatada por Lara (2009, p. 30), é o texto, e não o discurso. A opção justifica-se pelo fato de se encontrarem no texto "as estruturas informacionais, elementos que respondem pela coesão e progressão textual, bem como pelo seu fechamento e autonomia".

Desse modo, é no texto que se verifica a condição de registro, o que lhe dá características de permanência no tempo e de portabilidade no espaço. Essas mesmas condições são apontadas como requisito para que o que vai ser definido como informação por Smit e Barreto (2002).

A análise documentária é constituída por três atividades específicas: análise, síntese e representação (Kobashi, 1994). As duas primeiras fases dizem respeito à desestruturação do texto e a última à sua reestruturação.

Na etapa de análise propriamente dita é que ocorre a leitura documentária, momento em que o leitor profissional (indexador) realiza a leitura do texto e identifica o seu assunto. Neste momento, a concepção clara do que representa a ideia de texto, de suas tipologias e de como se configuram suas estruturas são recursos essenciais ao pleno êxito da operação.

Um texto, na definição funcional apresentada por Dias e Naves (2013, p. 27), é tomado como meio, na condição de

> veículo que permite a comunicação de ideias entre o sujeito que cria e dissemina informação (emissor, no caso autor) e o sujeito que necessita e adquire informação(receptor, no caso o leitor). É o objeto que permite a transmissão das informações contidas em documentos, sendo também visto como uma coleção de símbolos, os quais são intencionalmente estruturados pelo emissor para mudar a estrutura de imagem do receptor.

Na fase de análise, o leitor profissional precisa lançar mão de estratégias de leitura que possibilitem a compreensão do texto de modo completo, preciso e rápido. Para isso, utiliza-se, essencialmente, "de representações mentais, que estão ligadas ao conhecimento linguístico, social e de representação" (REDÍGOLO; FUJITA, 2015, p. 357). Um texto é

composto por uma rede interna e externa de relações. É essa rede que, entre outros elementos, irá viabilizar a construção de sentido pelo leitor.

A compreensão do texto condiciona-se às melhores condições de interação entre o texto, o leitor e o contexto de ocorrência da atividade de leitura. Desse modo, é preciso reafirmar o papel ativo do leitor durante o processo, mormente no processo de leitura profissional.

Cintra (1989, p. 34) distingue dois tipos de estratégias relativas ao processo de leitura: estratégias cognitivas e estratégias metacognitivas. As primeiras "compreendem comportamentos automáticos e inconscientes", as últimas "supõem comportamentos desautomatizados, na medida em que o leitor tem consciência de como está lendo". O acionamento dos conhecimentos prévios como estratégia metacognitiva de leitura, por exemplo, dá ao leitor profissional uma perspectiva ampliada de análise, compreensão e representação do texto nas tarefas de leitura documentária.

O conhecimento prévio, também chamado algumas vezes de "conhecimento de mundo", refere-se a todo o conhecimento que o indivíduo possui, isto é, o conhecimento que o indivíduo armazenou na mente como resultado da capacidade inata de organização das suas experiências com o mundo (MEURER, 1985), diz respeito à "memória semântica" e à "memória episódica" do indivíduo, ou seja, a todo o conhecimento generalizado e particularizado armazenado na mente.

O acionamento do conhecimento prévio, bem entendido, não se configura como um mecanismo de acionamento "liga-e-desliga". Está presente em qualquer atividade de leitura. O modo como isso ocorre, de forma consciente ou não é que irá determinar a natureza das estratégias metacognitivas ou cognitivas de leitura. A considerar-se (hipoteticamente) um nível zero de conhecimento prévio, a interpretação de um objeto de natureza complexa como o texto escrito torna-se, mais do que difícil, literalmente desprovida de sentido e impraticável.

A eficácia da leitura depende também, como se disse anteriormente, da qualidade do texto, naturalmente. Neste caso, há variáveis intervenientes que independem do leitor (como o estilo de redação e os recursos argumentativos do autor) e há as que são perfeitamente apreensíveis, como

ocorre com a percepção consciente da estrutura textual e dos recursos linguísticos utilizados.

O sentido geral que se procura alcançar com o texto é baseado na sua estrutura, a qual compreende suas micro, macro e superestrutura. Dias e Naves (2013, p. 29) apresentam as seguintes definições para estes termos: microestrutura: "estrutura superficial, que corresponde a realidade física do texto e seus símbolos de significação, as palavras"; macroestrutura: "concebida como um tópico representativo hierárquico e coerente da unidade textual, envolvendo mínima estrutura da representação textual, sintático-semântica" e superestrutura. "estrutura retórico-esquemática, um tipo de esquema de produção convencional para o qual o texto é adaptado, podendo ser considerado como transição entre estruturas de superfície e de profundidade".

Nesta perspectiva, quanto mais conhecimento a respeitos dos gêneros e das estruturas textuais o leitor profissional tiver, maiores serão suas chances de identificar com clareza a ização das informações consideradas essenciais no texto, ou seja, maiores as possibilidades de análise e compreensão do texto visando a sua representação para fins documentários.

Considerando-se uma abordagem *top-down*, compreende-se o documento, objeto da leitura documentária, como uma complexa rede de relações conceituais. Em sentido inverso, em abordagem *bottom-up*, observa-se o conceito como unidade mínima de significação na comunicação de conhecimento que tem o texto como veículo. Trata-se de um "jogo" por meio do qual autor codifica as relações entre os conceitos que pretende tratar no texto (percurso onomasiológico, do enunciador) e estabelece comunicação com o leitor, cuja tarefa é decodificá-las (percurso semasiológico, do interpretante).

Não se pode ler profissionalmente e de modo proficiente, portanto, sem o recurso de estratégias metacognitivas de leitura. A leitura é um processo de interação entre texto, leitor e contexto e requer, para sua plena efetivação, a presença de um leitor ativo, consciente da sua função na construção de sentidos, na reconexão entre os conceitos.

3 Os padrões de Hearst

Marti A. Hearst, professora e pesquisadora na *School of Information, University of California*, Berkeley e que tem dentre seus temas de pesquisa interesses na linguística computacional, propõe em alguns textos (HEARST, 1992; 1998) métodos para a identificação automática e a utilização de padrões léxico-sintáticos como recursos para a expressão de relações léxico-semânticas.

As pesquisas de Hearst tomam como base grandes corpora de textos que estão disponíveis na internet para extração das informações lexicais, sintáticas e semânticas. O método de "extração de padrões léxico-sintáticos" (*Lexico-Syntactic Pattern Extraction* - LSPE), conforme apresentado em Hearst (1998, p. 1, tradução livre), "pretende ser útil como uma ajuda automatizada ou semiautomatizada para lexicógrafos e construtores de bases de conhecimento dependentes de domínio".

Os padrões de Hearst, respeitando-se a terminologia que a pesquisadora prefere utilizar, possibilitam a identificação de relações conceituais de hiperonímia e de hiponímia que ocorrem por meio das posições que tais conceitos assumem relativamente uns aos outros.

Uma análise etimológica breve dos termos deixa ver claramente seus significados. O prefixo "hiper" diz respeito a algo que está em posição superior (ou em excesso), como ocorre, por exemplo, em "hipertensão" (def.: "Med. Pressão excessiva exercida pelo sangue nas paredes dos vasos sanguíneos"); o prefixo "hipo", por sua vez, relaciona-se a algo que está em posição inferior (ou em escassez), como ocorre, por exemplo, em "hipotensão" (def.: "Med. Pressão do sangue nas paredes dos vasos sanguíneos inferior à normal; pressão baixa").

Assim, hiperonímia é a palavra que transmite o sentido do todo e hiponímia remete à ideia de parte, tipo ou item do todo. Um hiperônimo é um termo que está superordenado em relação a um hipônimo; este, por sua vez, está subordinado em relação ao hipônimo. Trata-se, como se vê, de uma relação assimétrica. O termo "árvores frutíferas" é hiperônimo de "macieiras", "abacateiros" e "mangueiras". Tomando-se a relação em sentido oposto, "macieiras", "abacateiros" e "mangueiras" são hipônimos de "árvores frutíferas". Trata-se também de um tipo de relação estrutural, ao

mesmo tempo em que "árvores frutíferas" é hiperônimo de "macieiras" pode ser também hipônimo de "árvores".

Para que se possa compreender com mais clareza a natureza dos padrões utilizados por Hearst, pode-se observar um dos exemplos que a autora apresenta (HEARST, 1998, p. 3, tradução livre): "Agar é uma substância preparada a partir de uma mistura de algas vermelhas, tais como Gelidium, para uso laboratorial ou industrial".

A maioria dos leitores desta frase (considerando-se o domínio específico em que se inserem este texto que ora se desenvolve e seus leitores potenciais), desconhece o significado do termo "gelidium". O modo como estão estruturadas as informações na frase, entretanto, permite inferir que "gelidium"' é uma espécie de "algas vermelhas". Neste caso, não ocorre uma definição clássica, não se está deliberadamente definindo o termo. O que possibilita a construção semântica é a presença de um padrão léxico-sintático, nomeadamente o padrão "tais como" que conecta os conceitos "algas vermelhas" e "gelidium".

Observe-se outro exemplo com a aplicação do mesmo padrão: "A arquivística é tratada como a disciplina que agrupa todos os princípios, normas e técnicas que regem as *funções de gestão dos arquivos*, **tais como** a *criação*, a *avaliação*, a *aquisição*, a *classificação*, a *descrição*, a *comunicação* e a *conservação*" (GARCIA; SCHUCH JÚNIOR, 2002, p. 46, grifos acrescentados). Neste caso, a organização léxico-sintática das informações possibilita a identificação de relações hierárquicas entre os conceitos envolvidos que estão mediados pelo padrão "tais como". Estas relações, utilizando-se o modelo de visualização normalmente adotado nos tesauros, pode ser apresentada da seguinte forma:

---

funções de gestão de arquivos

TE criação

avaliação

aquisição

classificação

descrição

comunicação

conservação

Os padrões de Hearst, acredita-se, possuem potencial de aplicação na organização da leitura documentária. Tal potencial pode ser detectado nas características que Hearst (1998, p. 4, tradução livre) aponta relativamente ao conjunto de padrões léxico-sintáticos que indicam relações hierárquicas:

a) "eles ocorrem frequentemente em muitos gêneros textuais;

b) "eles (quase) sempre indicam a relação de interesse;

c) "eles podem ser reconhecidos com pouco ou nenhum conhecimento pré-codificado".

Os padrões identificados por Hearst (1998) em seu estudo relativo à língua inglesa, são descritos no Quadro 1.

Quadro 1 – *Os padrões de Hearst*

Padrão	**Exemplo**
NP_0 such as NP_1{,NP_2 ... , (and | or) NP_i} Obs.: NP = noun phrase ⇔ sintagma nomimal	*Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.* *-- agar* *---- gelidum*
such NP_0 as {NP_1,}* {(and | or)} NP_2	*... works by such authors as Herrick, Goldsmith, and Shakespeare.* -- author ---- Herrick ---- Goldsmith ---- Shakespeare
NP_1 {, NP_1}* {,} or other NP_0	*Bruises, ..., broken bones or other injuries ...* *-- injury* *---- broken bones* *---- bruises*

NP_1 {, NP_2}* {,} and other NP_0	*... temples, treasuries, and other important civic buildings.* *-- civic buildings* *---- temples* *---- treasuries*
NP_0 {,} including { NP_1 ,}* {or | and} NP_2	*All common-law countries, including Canada and England ...* *-- comom-law countries* *---- Canada* *---- England*
NP_0 {,} especially { NP_1 ,}* {or | and} NP_2	*... most European countries, especially France, England, and Spain.* *-- european countries* *---- France* *---- England* *---- Spain*

Fonte: *Elaborado pelos autores com base em Hearst (1992; 1998)*

Os padrões de Hearst foram identificados para aquisição automática de relações lexicais hiponímicas em língua inglesa. Considerando-se, contudo, seu potencial de aplicação também em outros idiomas, alguns estudos têm sido desenvolvidos para a identificação de padrões léxico-sintáticos que evidenciam relações léxico-semânticas também em língua portuguesa.

Machado e Lima (2015) descrevem em sua pesquisa alguns trabalhos correlatos, tanto com aplicação em língua estrangeira, principalmente o inglês, quanto em língua portuguesa. Os autores fazem referências enfáticas aos trabalhos de Hearst (1992, 1998) e apontam de modo mais genérico os autores que utilizaram os padrões de Hearst como base para aplicações em línguas estrangeiras, entre eles estão Maedche e Staab (2002), Cederberg e Widdows (2003) e Degeratu e Hatzivassiloglou (2004). Machado e Lima (2015) também indicam trabalhos em que aparecem traduções dos padrões de Hearst visando a sua aplicação em língua portuguesa, tais como Freitas (2007), Freitas e Quental (2007), Baségio (2007), Taba (2013), Taba e Caseli (2014) e Machado (2015).

Não foram identificados trabalhos relativos aos padrões de Hearst com aplicações no domínio da ciência da informação. Os trabalhos citados

anteriormente que discutem a aplicação dos padrões de Hearst em língua portuguesa são academicamente relacionados às áreas da linguística ou da ciência da computação. A pesquisa de Freitas (2007), *e.g.*, está vinculada ao Departamento de Letras da Pontifícia Universidade Católica do Rio de Janeiro; Freitas e Quental (2007) publicaram seu trabalho no *Workshop* de Tecnologias da Informação e da Linguagem Humana; Taba e Caseli (2014) publicaram sua pesquisa no *International Conference on Language Resources and Evaluatione* o trabalho de Machado e Lima (2015) foi publicado na Revista Estudos da Linguagem. Os trabalhos mais próximos da ciência da computação são: Taba (2013), que desenvolveu sua pesquisa no Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, Baségio (2007) e Machado (2015), cujas pesquisas vinculam-se ao Programa de Pós-graduação em Ciência da Computação da Pontifícia Universidade Católica - RS.

Freitas (2007) optou por utilizar três padrões de Hearst e identificou adicionalmente três outros padrões com observações do *corpus* utilizado em sua pesquisa, como pode ser observado na Quadro 2.

Baségio (2007) trabalhou com os padrões de Hearst (1992) e fez uma adaptação na expressão "sintagma nominal" (*noun phrase* – NP) para substantivo (SU). Isto foi necessário, justifica o autor, porque a composição do corpus não permitia dispor da informação que era necessária, deste modo simplificou-se o sintagma nominal. O autor supracitado utiliza outros padrões além dos que foram propostos por Hearst (1992) e outros autores citados, porém não compõem o interesse deste estudo por se referirem de modo específico à língua francesa.

Taba (2013) trabalhou com dois dos padrões de Hearst, a partir das adequações realizadas por Freitas (2007) e Freitas e Quental (2007), como pode ser observado no Quadro 2. Taba (2013) utilizou de modo manual em seu estudo dois padrões criados por Freitas (2007) e Freitas e Quental (2007) e identificou, para os fins de sua análise e de seu *corpus*, a necessidade de criar mais três padrões. Estes padrões estão relacionados com a extração de hiponímia, considerada pelos autores como resultante da relação semântica "is-a".

Machado (2015) refere-se aos padrões como "regras", e, como pode ser observado no Quadro 2, fez uso de cinco padrões de Hearst, observando as adaptações de Freitas (2007), Baségio (2007) e Taba (2013). A partir destes estudos Machado (2015) desenvolveu quatro "regras".

Quadro 2 – *Os padrões de Hearst e suas adaptações*

Padrão Hearst (1992, 1998)	**Freitas (2007) Freitas e Quental (2007)**	**Baségio (2007)**	**Taba (2013) Taba e Caseli (2014)**	**Machado (2015) Machado e Lima (2015)**
NP_0 such as NP_1{,NP_2 ... , (and | or) NP_i}	SN HHiper (tais como | como_PDEN) SN1 { , SN2 … , } (e | ou) SNi SN Hiper, (tais como | como_ PDEN) SN1 { , SN2 ... ,}(e | ou) SNi	SUB como {(SUB,)*(ou|e)} SUB SUB tal(is) como {(SUB,)*(ou|e)} SUB	SN_Hiper (tais como | como) SN {, SN} *(e|ou) SN	SN(,)? como (SN ,)*(SN (e|ou))*SN
such NP_0 as {NP_1,}* {(and | or)} NP_2	não utilizado	tal(is) SUB como {(SUB,)*(ou|e)} SUB	não utilizado	SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
NP_1 {, NP_1}* {,} or other NP_0	SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper	SUB {, SUB}* {,} ou outro(s) SUB	SN {, SN}* , ? (e|ou) outros SN_Hiper	SN (ou|e|,))*<outr(a|o)(s)? sn>
NP_1 {, NP_2}* {,} and other NP_0		SUB {, SUB}* {,} e outro(s) SUB		
NP_0 {,} including { NP_1 ,}* {or | and} NP_2	não utilizado	SUB {,} incluindo {SUB,}*{ou|e} SUB	não utilizado	SN(,)? incluindo (SN ,)*(SN (e|ou))*SN

NP_0 {,} especially { NP_1 ,}* {or | and} NP_2	não utilizado	SUB {,} especialmente {SUB,}*{ou|e} SUB SUB {,} principalmente {SUB,}*{ou|e} SUB SUB {,} particularmente {SUB,}*{ou|e} SUB SUB {,} em especial { SUB,}*{ou|e} SUB SUB {,} em particular { SUB,}*{ou|e} SUB SUB {,} de maneira especial { SUB,}*{ou|e} SUB SUB {,} sobretudo { SUB,}*{ou|e} SUB	não utilizado	SN(,)? especialmente (SN ,)*(SN (e|ou))*SN
-	tipos de SN Hiper: SN1{ , SN2... ,} (e | ou) SNi	-	tipos de SN_ Hiper: SN {, SN}* (e|ou) SN	<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN
-	SN HHiper *chamado/s/a/as*(de) SN Hipo	-	SN_Hiper chamad(o|a|os|as) de? SN	SN(,| é| são| foram)? chamad (o|a|os|as)(de)? (SN ,)* (SN (e|ou))*SN

-	*SN Hiper *conhecido/s/a/as como* SN Hipo.	-	-	SN((,)? também)? (,|é|são|foram)? conhecid (o|a|os|as) como (SN ,)*SN (e|ou))*SN"
-	-	-	SN {,SN} * ,? (e|ou) (qualquer| quaisquer) outro{s}? SN_ Hiper	(SN (ou|e|,))* < (qualquer| quaisquer) outr(a|o)(s)? sn>
-	-	-	SN é (o|a|um|uma) SN_Hiper	SN é < (o|a) sn> SN é < (um|uma) sn>
-	-	-	SN são SN_ Hiper	SN são SN

Fonte: *Elaborado pelos autores*

Cabe salientar que as pesquisas desenvolvidas levaram em consideração, como era de se esperar, os trabalhos previamente publicados. Desse modo, Machado (2015) utilizou o artigo da Taba e Caseli (2014) e Taba (2013) utilizou o artigo de Freitas e Quental (2007). Tendo-se isto em mente, neste estudo que se apresenta foram observados mais proximamente os trabalhos dos autores tomados como "seminais": Freitas (2007), Baségio (2007), Taba (2013) e Machado (2015). Foram estes trabalhos que serviram de base para o desenvolvimento da fase experimental desta pesquisa e que subsidiaram as adaptações necessárias à análise.

A literatura atesta a potencialidade dos padrões de Hearst para a identificação de relações hierárquicas no texto. Esta literatura forma, contudo, um grande mosaico a respeito das questões particulares a cada contexto de análise. Os trabalhos realizados por Freitas (2007), Baségio (2007), Taba (2013) e Machado (2015), foram desenvolvidos a partir de *corpus* já analisados e etiquetados sintaticamente. Baségio (2007), Taba (2013) e Machado (2015) desenvolveram em suas pesquisas um protótipo de *software*. Em Taba (2013) este software é denominado "Anotador de Relações Semânticas (ARS)".

Para fins de verificação de aplicabilidade dos padrões de Hearst no contexto da leitura documentária, observando-se de modo específico a identificação de termos e suas relações hierárquicas de modo semiautomático, foi realizado um experimento cujos procedimentos estão descritos na seção subsequente.

4 Procedimentos metodológicos

O experimento realizado neste estudo foi desenvolvido a partir de um *corpus* selecionado da Revista Latino-Americana de Enfermagem (RLAE) da Escola de Enfermagem de Ribeirão Preto - Universidade de São Paulo. Utilizou-se o seu volume mais recente (v. 25), contendo 26 artigos e um editorial. A escolha do referido periódico deu-se de modo aleatório, mas considerou-se sua qualificação no estrato "Qualis/Capes A1" da área de saúde, mais especificamente em enfermagem. Conforme informações disponíveis na página da RLAE, trata-se de um periódico bimestral que circula desde janeiro de 1993.

Para a realização da análise, foram executados os procedimentos de *download* dos textos em formato PDF e em língua portuguesa (a revista disponibiliza também versões em língua espanhola e inglesa dos artigos). A ferramenta escolhida para análise do corpus foi o *softwareWordSmith Tools* versão 4.0, com a licença *free*. Executou-se o procedimento de conversão para arquivos no formato TXT, legível pelo *WordSmith*, e realizou-se o tratamento manual dos textos para evitar a separação de parágrafos, já que acontece, muitas vezes, que na conversão de arquivos alguns parágrafos fiquem separados por quebras de páginas, figuras, tabelas, entre outros.

O *WordSmith Tools* é um conjunto integrado de funções para análise do comportamento das palavras nos textos; as principais funções são: *WordList*, *Concord* e *KeyWords*. O *WordList* possibilita visualizar uma lista de palavras ou um agrupamento. A função *Concord*, permite a ização de termos em textos e apresenta o contexto que aquele termo aparece, ou seja,possibilita listar as ocorrências e consultá-las na íntegra. O *KeyWords*possibilita a ização de um conjunto de palavras-chave nos textos. Além destas funções principais é possível obter outras informações estatísticas dos textos, como padrões, frequências, entre outras.

O *WordSmith Tools* foi desenvolvido por Mike Scott na *Oxford University Press* para o desenvolvimento de seu próprio trabalho lexicográfico e atende a diversos idiomas, inclusive o português.

Nesta pesquisa, trabalhou-se com duas modalidades de análise: a) semiautomática, a partir do uso da ferramenta *WordSmith* e b) manual. Na análise semiautomática foi utilizado o *corpus* em sua totalidade. Esta opção foi possível por conta do retorno rápido e preciso que a ferramenta tecnológica proporciona. Já na análise manual, foram utilizados os três primeiros artigos da revista.

A opção pela aplicação dos padrões de Hearst (1992, 1998) neste estudo foi fundamentada, como já se disse anteriormente, nos resultados dos trabalhos desenvolvidos por Freitas (2007), Baségio (2007), Taba (2013) e Machado (2015). Assim, foram utilizados os padrões descritos no Quadro 3, já com as devidas adaptações para efeitos de simplificação quanto ao seu entendimento e aplicação, mas mantendo-se, de qualquer modo, sua relação com os padrões de Hearst (1992, 1998). Como ocorre em quase toda simplificação, há, neste caso, alguma redução em relação ao algoritmo apresentado nos padrões originais.

Quadro 3 – *Os padrões léxico-sintáticos aplicados na pesquisa*

PADRÕES DE HEARST (1992, 1998)	PADRÕES APLICADOS
NP_0 such as NP_1{,NP_2 ... , (and | or) NP_i}	(1) SN (tais como | como) SN { , SN … , } (e | ou) SN
such NP_0 as {NP_1,}* {(and | or)} NP_2	(2) tal(is) SN como {(SN,)*(ou|e)} SN
NP_1 {, NP_1}* {,} or other NP_0	(3) SN {, SN}* {,} ou outro(s) SN
NP_1 {, NP_2}* {,} and other NP_0	(4) SN {, SN}* {,} e outro(s) SN
NP_0 {,} including { NP_1 ,}* {or | and} NP_2	(5) SN {,} incluindo {SN,}*{ou|e} SN
NP_0 {,} especially { NP_1 ,}* {or | and} NP_2	(6a) SN {,} especialmente {SN,}*{ou|e} SN (6b) SN {,} principalmente {SN,}*{ou|e} SN (6c) SN {,} particularmente {SN,}*{ou|e} SN (6d) SN {,} em especial { SN,}*{ou|e} SN (6e) SN {,} em particular { SN,}*{ou|e} SN (6f) SN {,} de maneira especial { SN,}*{ou|e} SN (6g) SN {,} sobretudo { SN,}*{ou|e} SN

Fonte: *Elaborado pelos autores*

Em relação aos procedimentos de aplicação, o algoritmo de Hearst (1992), é orientado pelos seguintes passos: 1) escolher a relação semântica; 2) obter uma lista de termos para validar as relações; 3) izar no corpus onde essas relações ocorrem sintaticamente próximas e registrar as ocorrências; 4) izar similaridades entre esses registros e a possível indicação da relação de interesse; 5) identificando-se positivamente um novo padrão, utilizá-lo para coletar mais instâncias da relação alvo e retornar à etapa 2.

Em relação ao passo 1, foram definidas as relações semânticas de hiperonímia e de hiponímia. Para a consecução do passo 2, utilizou-se a ferramenta *parserWordSmith*, anteriormente descrita, para izar no próprio *corpus* as relações que foram tomadas como relações positivas. Considerando-se que a lista de relações e seus termos foram extraídos do próprio *corpus* por meio da ização do termo-chave de cada padrão, os passos 3 e 4 foram executados juntamente com o passo 2.

Foram esgotadas as buscas por meio do termo-chave de cada um dos padrões e as ocorrências foram analisadas de maneira manual, homologando-se as hierarquias encontradas. A partir das identificações das relações, foram realizadas as análises visando ao atendimento do passo 5, isto é, avaliar positivamente as relações hierárquicas.

5 Identificação de estruturas hierárquicas com recurso dos padrões de Hearst

Os resultados estão estruturados tomando-se os padrões como categorias de análise. Apresentam-se, para cada padrão, excertos do *corpus*com a finalidade de ilustrar a sua aplicação. Apresentam-se adicionalmente, para os casos possíveis, a estrutura hierárquica conforme poderia ocorrer sua configuração em um tesauro.

Os exemplos apresentados variam em conformidade com o interesse da exposição dos resultados. Assim, tanto são apontados e discutidos casos em que os padrões permitiram a identificação de relações hierárquicas como casos em que isso não foi possível.

Na categorização das relações extraídas, verificou-se também algumas situações de atendimento parcial dos padrões, ou seja, casos em que

as ocorrências apresentavam algum conectivo não previsto entremeando os sintagmas nominais e os padrões.

Houve também casos em que não foi possível identificar a qualidade e a identidade lógica da relação hierárquica pela ausência de sentido geral no contexto do corpus analisado. Estes casos demandavam pesquisas como forma de homologar a relação hierárquica que não estão entre os objetivos deste trabalho.

Os resultados estatísticos gerais relativos às aplicações dos padrões de Hearst e seus índices de ocorrência no *corpus* podem ser verificados na Tabela 1, apresentada após as subseções que apresentam e discutem os padrões.

5.1 Padrão "tais como" / "como"

O padrão originalmente proposto por Hearst, adaptado e utilizado por Freitas (2007), Baségio (2007), Taba (2013) e Machado (2015), foi utilizado neste estudo. Este padrão é representado da seguinte forma:

- SN (tais como | como) SN { , SN … , } (e | ou) SN.

Aplicou-se parte do padrão "tais como" para recuperar os registros de ocorrência por meio da ferramenta *WordSmith*, obtendo-se como resultado 23 ocorrências no *corpus*. Posteriormente aplicou-se a consulta utilizando-se "como". Considerando-se que a relação lógica entre os componentes "tais como" e "como" neste padrão se dão pela presença de um "ou" exclusivo e observando-se que os resultados alcançados com o uso do "como" foram praticamente todos contemplados com o uso de "tais como", procedeu-se a exclusão dos registros recuperados em duplicidade, resultando, no total, 494 ocorrências.

Apresentam-se, na sequência, exemplos de aplicação do padrão "tais como | como", os quais podem ser visualizados por meio dos excertos 1 e 2 e de suas estruturas hierárquicas correspondentes. Estes dois exemplos, são considerados como positivos pelo sucesso na sua aplicação, isto é, pela extração de relação hierárquica pode meio da identificação dos

padrões léxico-sintáticos. O hiperônimo está izado no primeiro sintagma nominal antes da identificação do padrão (destacado em negrito) e os hipônimos logo depois.

Excerto 1 - "Mas alguns trabalhos rotineiros têm duplicações com os cuidados hospitalares, **tais como** infusões intravenosas, injeções e curativos".

---

cuidados hospitalares

curativos

infusões intravenosas

injeções

Excerto 2 - "[...] Na atenção primária e em centros de atendimento ambulatorial, a atenção com a NP para pacientes com doenças crônicas, **como** doenças cardíacas, hipertensão e diabetes, resultou em melhores indicadores de controle das doenças [...]".

---

doenças crônicas

diabetes

doenças cardíacas

hipertensão

O excerto 3 é uma amostra de caso em que o padrão não se aplica por não se verificar a ocorrência de sintagmas nominais, mas sim a presença de sintagma verbal antes do "como" previsto no padrão. Na verdade, a ocorrência "como", neste caso, tem aplicação completamente diversa do que prevê o padrão "como".

Excerto 3 - "[...] Seria mais apropriado empregar **como** denominador uma estimativa da exposição à gravidez de meninas de 10 a 14 anos de idade, no entanto estes dados não estavam disponíveis. [...]".

Vale observar que, em relação ao padrão "tais como", não foi verificada nenhuma ocorrência que impossibilitasse um enquadramento positivo, ainda que os aspectos semânticos das relações hierárquicas não sejam facilmente visíveis de modo automático. O excerto 4 ilustra este caso.

Excerto 4 - "[...] Na Parceria de Apoio, os profissionais de saúde apoiam os membros da comunidade conforme necessário de acordo com suas situações, **tais como** doenças ou condições de envelhecimento".

---

situações de apoio da Parceria de Apoio

condições de envelhecimento

doenças

5.2 Padrão "tal(is)"

O padrão de Hearst denominado "tal(is)" não foi utilizado por Freitas (2007) e Taba (2013), mas aparece em Baségio (2007) e Machado (2015). Este padrão está representado como segue:

- tal(is) SN como {(SN,)*(ou|e)} SN.

O padrão "tal(is)" foi izado em apenas duas ocorrências. Em ambas não foi possível a extração positiva das relações hierárquicas, por suas inadequações em relação ao padrão. No excerto 5 é possível observar a satisfação parcial do que contempla o padrão.

Excerto 5 - "[...] Em tal contexto, o enfermeiro, **como** profissional de saúde, tem papel fundamental na elaboração e prática de intervenções que modifiquem essa realidade".

No excerto 6 há uma distância considerável entre "tal" e "como", o que inviabiliza o estabelecimento seguro de aplicação do padrão.

Excerto 6 - "[...] **Tal** categoria apreendeu conteúdo para cinco subcategorias que, no referido programa de computador, são traduzidas **como** codes".

5.3 Padrão "ou outro(s)"

Este padrão foi utilizado por Freitas (2007), que incorporou a este padrão o padrão "e outro(s)" (objeto da próxima subseção, de número 5.4). Do mesmo modo fizeram Taba (2013) e Machado (2015). Baségio (2007) tratou separadamente os dois padrões, do mesmo modo como ocorreu em relação a este trabalho de pesquisa. Este padrão apresenta-se do seguinte modo:

- SN {, SN}* {,} ou outro(s) SN.

Para o padrão "ou outro(s)", foram obtidas cinco ocorrências no texto, destas somente uma não compreende de modo positivo o padrão, demonstrada no excerto 7. Este exemplo reforça o argumento de que não se pretende, a partir do que é traçado nesta pesquisa, eleger abordagens completamente automáticas de leitura.

Excerto 7 - "[...] ou seleção aleatória, para os grupos, todas as participantes tiveram uma chance igual de serem incluídos em um **ou outro** grupo".

As demais ocorrências todas atendem ao padrão. No excerto 8, que ilustra um caso positivo, é possível observar que o hiperônimo está logo após o "ou outro(s)" e o hipônimo antes. Este exemplo ilustra também que não se pode dispensar, em qualquer caso, o recurso aos instrumentos de controle de vocabulário, como os tesauros. Neste caso, o próprio texto faz menção à substituição da expressão "líquidos corporais" por "fluidos corporais".

Excerto 8 - "[...] Eu limpo imediatamente com desinfetante (álcool) superfícies após derramamento de sangue **ou outros** líquidos corporais.

---

fluidos corporais

sangue

5.4 Padrão "e outro(s)"

Este padrão é representado como segue:

- SN {, SN}* {,} e outro(s) SN.

No total, foram izadas no corpus, 22 ocorrências relativas a este padrão. Destas, duas não atenderam, três atenderam parcialmente e 17 atenderam de modo positivo a aplicação do padrão em destaque.

Apresenta-se no excerto 9 um dos casos em que a aplicação do padrão ocorreu de modo positivo.

Excerto 9 - "[...] É impossível planejar a APS sem ter claros os papéis para os médicos, enfermeiras, parteiras **e outros** profissionais de saúde.

---

profissionais da saúde

enfermeiros

médicos

parteiros

Pode-se observar que hiperônimo fica izado após a ocorrência do padrão "e outro(s)" e os hipônimos antes dele, de modo semelhante ao que ocorreu com o padrão "ou outro(s)".

O excerto 10 demonstra o não atendimento ao padrão. A expressão "e outro", neste caso, é recurso estilístico para diferenciar os dois tipos de agentes envolvidos no que se relata, sem que apresentem relações de hierarquia entre si.

Excerto 10 - "[...] Além delas, foram escolhidos dois profissionais – um que realizou as intervenções comportamentais **e outro** que realizou as intervenções educativas por telefone".

5.5 Padrão "incluindo"

Este padrão não foi utilizado por Freitas (2007) e Taba (2013). Baségio (2007) e Machado (2015), entretanto, aplicaram-no em suas pesquisas. Eis sua representação:

- SN {,} incluindo {SN,}*{ou|e} SN.

Foram identificadas 57 ocorrências deste padrão no *corpus*. Em trinta delas, houve total compatibilidade com o que se esperava da aplicação do padrão, isto é, permitiram a identificação de hiperônimos e hipônimos. Em dezenove ocorrências houve resposta parcial e em oito casos a expressão identificada guardava apenas relação de similaridade com o padrão, ou seja, não havia relações hierárquicas entre os termos que a expressão tomada como padrão mediava.

No excerto 11 é possível observar a relação hierárquica identificada com a aplicação do padrão "incluindo". Identificam-se o hiperônimo como seu antecessor e os hipônimos como sucessores.

Excerto 11 - "[...] Resultados: participaram do estudo 573 profissionais, **incluindo** técnicos e auxiliares de enfermagem 292 (51%), enfermeiros 105 (18,3%), médicos 59 (10,3%), e outros profissionais 117 (20,4%)".

---

profissionais

auxiliares de enfermagem

enfermeiros

médicos

técnicos de enfermagem

Cabe mais uma vez a advertência: os casos são apresentados como exemplos e procurou-se manter o máximo possível de fidelidade em re-

lação ao contexto de ocorrência para que fosse possível compreender a aplicação dos padrões de Hearst. O termo "profissionais" no excerto 11, por exemplo, mantido na estrutura hierárquica apresentada, é vago. Além do próprio contexto de ocorrência do padrão (uma revista de enfermagem) informar, há menção explícita no texto aos "profissionais da saúde". Independentemente disto, contudo, a utilização de um tesauro como recurso para a tradução no processo de indexação resolveria, pela sugestão de termos mais precisos, a questão.

O exemplo apresentado no excerto 12 é um dos casos em que a relação hierárquica não pôde ser identificada plenamente e de modo formal pela aplicação automática do padrão "incluindo". Um leitor profissional, ou mesmo um outro tipo de leitor proficiente, percebe facilmente que as expressões que aparecem depois da palavra "incluindo" ("utilização de EPI", "uso de sistemas de exaustão " e "ventilação eficaz nas salas operatórias") são relativas às "medidas preventivas nos CC para a minimização dos riscos químicos devido à exposição à inalação da fumaça cirúrgica", funcionando com seus hipônimos. Ocorre que a expressão é muito longa para funcionar de modo adequado como descritor e nem mesmo a utilização pura e simples de um vocabulário controlado pode resolver prontamente a questão.

Excerto 12 - "[...] Portanto, tem-se como prioridade a adoção de medidas preventivas nos CC para a minimização dos riscos químicos devido à exposição à inalação da fumaça cirúrgica, incluindo a utilização de EPI e o uso de sistemas de exaustão e de ventilação eficaz nas salas operatórias".

O que se busca identificar nesta pesquisa, rememora-se, é a identificação de relações semânticas hierárquicas de modo semiautomático com o recurso de padrões léxico-sintáticos. Neste caso específico em análise, a resposta não é completamente satisfatória, justamente porque a tarefa maior de identificação dos sentidos atribuídos aos termos fica por conta do leitor, o que torna a utilização do padrão com recurso de leitura no mínimo discutível.

5.6 Padrão "especialmente" e semelhantes

Este padrão foi utilizado por Baségio (2007) e Machado (2015). Freitas (2007) e Taba (2013) não o aplicaram em suas pesquisas. Machado (2015) fez uso apenas do padrão "especialmente", Baségio (2007) ramificou este padrão para "principalmente", "particularmente", "em especial", "em particular", "de maneira especial" e "sobretudo". Para os efeitos desta investigação, optou-se pela aplicação e análise de todos estes padrões.

Como resultado geral da análise, verificou-se que os padrões "especialmente", "principalmente", "particularmente", "em especial", "em particular", "de maneira especial" e "sobretudo" são menos precisos, de modo geral, que os demais. No interior deste grupo, apenas o padrão "especialmente" destaca-se dos demais.

O índice de ocorrências positivas para este grupo de padrões foi de 29 acertos. Destes, o padrão "especialmente" responde por vinte acertos (em 39 ocorrências) e o padrão "principalmente" por quatro acertos (em 22 ocorrências). Considerando-se isto, optou-se por descrever apenas os resultados destes últimos padrões.

O padrão "especialmente" (e suas variações), pode ser representado do seguinte modo:

- SN {,} especialmente {SN,}*{ou|e} SN.

Como resultado da aplicação do padrão "especialmente", foram identificadas 39 ocorrências, sendo que vinte satisfazem totalmente ao que se espera do padrão, onze atendem parcialmente e oito não atendem, isto é, não possibilitam a extração de relações hierárquicas.

No excerto 13 é possível observar o hiperônimo, que antecede ao padrão "especialmente", e seus respectivos hipônimos.

Excerto 13 - "[...] São necessários mais serviços de promoção da saúde, prevenção e gestão para reduzir a carga de doença e a mortalidade associada a doenças crônicas, **especialmente** a saúde mental, câncer, doenças cardiovasculares e diabetes".

---

doenças crônicas

câncer

diabetes

doenças cardiovasculares

saúde mental

Neste caso, embora a aplicação e o resultado obtido com o padrão estejam qualificados como positivos, percebe-se claramente um erro lógico causado pela ambiguidade da redação do texto original: "saúde mental" não é, evidentemente, hipônimo de "doenças crônicas", talvez a ausência desta o seja. Conforme já dito anteriormente, procurou-se ater de modo específico ao corpus justamente para que se pudesse observar o comportamento na aplicação dos padrões de Hearst. Além disso, é importante lembrar que a qualidade da relação lógica não pode ser depreendida de um ou outro excerto, de um ou outro *corpus*, mas sim de uma verificação sistemática deles.

O excerto 14, reproduzido na sequência, ilustra um caso em que não há resposta satisfatória com a aplicação do padrão. A presença de sintagmas verbais antecedendo o padrão "especialmente" dificulta, neste caso, a extração confiável das relações hierárquicas.

Excerto 14 - "[...] Para suprimir as restrições financeiras de cuidados com a saúde, é importante melhorar o estilo de vida para prevenir as DNT e promover saúde, reforçando **especialmente** o conhecimento e alfabetização em saúde para todas as idades.

Em relação ao padrão "principalmente", uma derivação ou variação do padrão "especialmente", foram identificadas 22 ocorrências no *corpus*: quatro positivas, dez parciais e oito em que a identificação do padrão não apresentou resposta satisfatória.

No excerto 15 reproduz-se um caso positivo em que se verifica uma relação meronímica.

Excerto 15 - "[...] Desse modo, torna-se necessário o desenvolvimento de pesquisas para que se possam obter melhores conhecimentos sobre o ambiente de trabalho, **principalmente** nos serviços de emergência e como o mesmo interfere na prática profissional.

---

ambientes de trabalho

serviços de emergência

Na Tabela 1, que apresenta os resultados estatísticos gerais relativos às aplicações dos padrões de Hearst e seus índices de ocorrência no *corpus, foram*consideradas ocorrências todas as instâncias das expressões indicativas do padrão.

Tabela 1–*Distribuição geral dos padrões no corpus*

Nº	PADRÃO	OCORRÊNCIAS	RELAÇÕES	%	ACERTOS	%
1	SN (tais como | como) SN { , SN … , } (e | ou) SN	518	228	44,02	130	25,1
2	tal(is) SN como {(SN,)*(ou|e)} SN	2	1	50	-	-
3	SN {, SN}* {,} ou outro(s) SN	5	4	80	4	80
4	SN {, SN}* {,} e outro(s) SN	22	20	90,91	17	77,27
5	SN {,} incluindo {SN,}*{ou|e} SN	57	49	85,96	30	52,63
6a	SN {,} especialmente {SN,}*{ou|e} SN	39	31	79,49	20	51,28
6b	SN {,} principalmente {SN,}*{ou|e} SN	22	14	63,64	4	18,18

6c	SN {,} particularmente {SN,}*{ou|e} SN	2	2	100	1	50
6d	SN {,} em especial { SN,}*{ou|e} SN	1	1	100	1	100
6e	SN {,} em particular { SN,}*{ou|e} SN	6	2	33,33	-	-
6f	SN {,} de maneira especial { SN,}*{ou|e} SN	0	-	-	-	-
6g	SN {,} sobretudo { SN,}*{ou|e} SN	10	7	70	3	30
	TOTAL	**684**	**359**	**52,49**	**210**	**30,7**

Fonte: *Elaborada pelos autores*

A coluna "relações" refere-se às quantidades de relações positivas extraídas, incluindo-se as que atenderam parcialmente ao padrão. Considerou-se "acertos" todas as relações positivas, isto é, casos em que foi possível identificar relações semânticas a partir da identificação de relações lexicais.

Considerações finais

Pode-se considerar que há dois macro-processos orientados ao tratamento da informação encontrada em conjuntos textuais: a extração e a abstração. O primeiro relaciona-se aos padrões léxico-sintáticos, refere-se ao que está presente no texto, o segundo diz respeito ao que não está formalizado no texto. Os computadores, sabe-se, são mais eficientes na primeira tarefa, isto é, produzem resultados mais confiáveis quando aplicados em atividades de extração do que em atividades de abstração.

A atividade de leitura documentária para fins de indexação é, pelo conjunto de variáveis que envolve (já apresentadas anteriormente), extremamente complexa. Trata-se de uma atividade que demanda algum nível de extração, mas cuja natureza configura-se como essencialmente abstrata. Indexa-se não o que está no texto, ou pelo menos não necessariamente o que está no texto, mas sim o que não está nele. O que se quer identificar e representar com processo de leitura documentária são os conceitos; o acesso aos conceitos (o que está ausente), entretanto, é de

natureza representacional, intermediado pelos termos (o que está presente) nos textos científicos.

Os aspectos teóricos e os resultados apontados neste trabalho apontam para dois aspectos distintos e inter-relacionados referentes à leitura documentária: os quase insondáveis e complexos aspectos sociocognitivos que a envolvem, e formam o seu núcleo, e alguns aspectos formais que, a partir do conhecimento da estrutura textual do documento que se analisa e do recurso a algumas ferramentas adequadas, podem ser identificados a partir de determinados padrões léxico-sintáticos que correspondem a determinados "padrões" léxico-semânticos.

É preciso reafirmar que, conforme aponta este estudo e também os que foram apontados ao longo do trabalho, não existem, a rigor, padrões universais (nem se pretendia encontrá-los, aliás). Para cada cultura, para cada domínio, para cada corpus estabelece-se um padrão próprio que pode, como é próprio dos artefatos culturais, alterar-se. A noção de padrão, portanto, deve ser formada a partir de análises periódicas sistemáticas, originadas de *corpus* representativos dos domínios que se pretende representar.

Dada esta complexidade, nem sempre os padrões funcionam, ou melhor, nem sempre as expressões tomadas como padrão são empregadas da mesma forma. A língua, mesmo nas terminologias, não é apenas código. Por este motivo, não se insinua, em nenhum momento, o recurso aos padrões léxico-sintáticos como substituto ao leitor profissional, o que seria, nestas condições, ingenuidade, para dizer o mínimo. Compreendeu-se aplicação de tais recursos como relativos a uma forma de abordagem semiautomática, que visa a instrumentalizar o processo de leitura documentária.

Os usos dos padrões léxico-sintáticos indicados no estudo não estão livres das ambiguidades e das alterações de sentido que podem sofrer em função dos recursos estilísticos de quem os emprega. O padrão "tais como", por exemplo, foi identificado nos resultados como um dos que mais positivamente responderam à identificação de relações hierárquicas às quais entremeava. A expressão "tais como", entretanto pode ser perfeitamente empregada para indicar relações de outra natureza, como – para empregar um último exemplo – a que se pode observar na citação extraída

de Assumpção (2011, p. 95, grifos acrescentados): "Na literatura é possível encontrar algumas variaçõesterminológicas para denominar um arquivo de autoridade, tais como 'catálogo de autoridades', 'lista de cabeçalhos autorizados', 'lista de autoridades', 'catálogo de identidade', 'catálogo de formas autorizadas', entre outros". O leitor profissional, atento e com algum conhecimento acerca do domínio, percebe que não são de natureza hierárquica as relações entre os termos envolvidos. Essas relações referem-se a relações de equivalência: "catálogos de autoridade", "lista de cabeçalhos autorizados" etc. não são tipos-de ou parte-de "arquivos de autoridade", são antes variações terminológicas (ou sinonímicas), como está, aliás, explicitamente indicado no texto.

Acredita-se, por fim, que o acréscimo contínuo de textos em formato eletrônico e o incremento do volume de hipertextos não apenas requerem como possibilitam, em função de seus formatos, estruturas e suportes, o tratamento semiautomático da identificação de informações textualmente e contextualmente importantes. Às habilidades tradicionalmente requeridas ao bibliotecário relativas às operações da análise documentária devem ser acrescentados conhecimentos oriundos das ciências cognitivas, da linguística textual e da análise do discurso, entre outras.

Este estudo limitou-se à análise da aplicação dos padrões de Hearst a um *corpus* relativamente pequeno, visando testar a viabilidade de aplicação dos padrões léxico-sintáticos para identificação de relações léxico-semânticas. O volume do *corpus* não lhe dá caráter de representatividade em relação á área da enfermagem.

Conclui-se que a aplicação de dos padrões léxico-sintáticos selecionados para o estudo é potencialmente útil como recurso semiautomático para a realização da leitura documentária. Em estudos futuros, pretende-se ampliar o *corpus*incluindo outros domínios e o conjunto de padrões para que possam incluir, além das relações de hiperonímia e hiponímia, outras relações lexicais, como a meronímia e a holonímia, as relações de equivalência e as sempre mais complexas relações associativas.

Referências

ASSUMPÇÃO, F. A. *Controle de autoridade*: definições, processos e componentes. Monografia (Bacharelado em Biblioteconomia) - Universidade Estadual Paulista, Marília, 2011.

BASÉGIO, T. L. *Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil*.Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2007.

BOCCATO, V. R. C. *Avaliação do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias*: um estudo sociocognitivo com protocolo verbal. 2009. 301 f. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Marília, 2009.

CARNEIRO, M. V. Diretrizes para uma política de indexação. *Revista da Escola de Biblioteconomia da UFMG*, v. 14, n. 2, p. 221-241, set. 1985.

CEDERBERG, S.; WIDDOWS, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 7, 2003, Edmonton. *Proceedings...* Edmonton: Association for Computational Linguistics, 2003. CONLL, v. 4 p. 111-118.

CINTRA, A. M. M. Estratégias de leitura em documentação. In: SMIT, J. W. (Coord.). *Análise documentária*: a análise da síntese. 2.ed. Brasília: IBICT, 1989. p. 63-87.

DEGERATU, M.; HATZIVASSILOGLOU, V. An automatic method for constructing domain-specific ontology resources. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 4, 2004, Lisboa. *Proceedings...* Lisboa, 2004. p. 2001-2004.

DIAS, E. W.; NAVES, M. M. L. *Análise de assunto*: teoria e prática. 2.ed.rev. Brasília: Briquet de Lemos, 2013.

FREITAS, M. C.; QUENTAL, V. S. D. B. Subsídios para a elaboração automática de taxonomias. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 27.; Workshop em Tecnologia da Informação e da Linguagem Humana, 5., 2007, Rio de Janeiro. *Anais...* Rio de Janeiro: SBC, 2007.

FREITAS, M. C. *Elaboração automática de ontologias de domínio*: discussão e resultados.Tese (Doutorado em Letras). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

FUJITA, M. S. L. (Org.). *Política de indexação para bibliotecas*: elaboração, avaliação e implantação. Marília: Oficina Universitária, 2016.

FUJITA, M. S. L. *A leitura documentária do indexador*: aspectos cognitivos e linguísticos influentes na formação do leitor profissional. Tese (Livre-docência) - Universidade Estadual Paulista Júlio de Mesquita Filho, 2003.

FUJITA, M. S. L.; RUBI, M. P. Um modelo de leitura documentária para a indexação de artigos científicos: princípios de elaboração e uso para a formação de indexadores. *DataGramaZero*, v. 7, n. 3, jun. 2006.

GARCIA, O. M. C.; SCHUCH JÚNIOR, V. F. A aplicação da arquivística integrada, considerando os desdobramentos do processo a partir da classificação. *Informação & Informação*, v. 7, n. 1, p. 41-56, jan./jun. 2002.

GIL LEIVA, I.; FUJITA, M. S. L. (Eds.). *Política de indexação*. Marília: Oficina Universitária, 2012.

GUIMARÃES, J. A. C. *Análise documentária em jurisprudência*: elementos para uma metodologia de indexação de acórdãos trabalhistas brasileiros. Tese (Doutorado em Ciências da Comunicação) - Universidade de São Paulo, 1994.

HEARST, M. A. Automated discovery of WordNet relations. In: FELLBAUM, C. (Ed.). *WordNet*: an electronic lexical database and some of its applications. Cambridge: MIT Press, 1998.

HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14., Nantes, 1992. *Proceedings…* Nantes, 1992.

KOBASHI, N. Y**.** *A elaboração de informações documentárias*: em busca de uma metodologia. Tese (Doutorado em Ciências da Comunicação) – Universidade de São Paulo, São Paulo, 1994.

LANCASTER, F. W. *El control del vocabulario en la recuperación de información*. 2 ed. València: Universitat de València, 2002.

LARA, M. L. G. *A representação documentária*: em jogo a significação. Dissertação (Mestrado em Ciências da Comunicação) – Universidade de São Paulo, São Paulo, 1993.

LARA, M. L. G. *Linguística documentária*: seleção de conceitos. Tese (Livre-docência) – Universidade de São Paulo, São Paulo, 2009.

LARA, M. L. G. *Representação e linguagens documentárias*: bases teórico-metodológicas. Tese (Doutorado em Ciências da Comunicação) – Universidade de São Paulo, São Paulo, 1999.

LUCAS, C. R. *Leitura e interpretação em Biblioteconomia*. Campinas: Unicamp, 2000.

MACHADO, P. N.; LIMA, V. L. S.. Extração de relações hiponímicas em um corpus de língua portuguesa. *Revista de Estudos da Linguagem*, v. 23, n. 3, p. 599-640, dez. 2015.

MACHADO, P. N. *Extração de relações hiponímicas em corpora de língua portuguesa*. Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2015.

MAEDCHE, A.; STAAB, S. *Ontology learning for the semantic web*. Massachusetts: Kluwer Academic, 2002.

MEURER, J. L. Schemata in reading comprehension. *Ilha do Desterro*, n. 13, p. 31-46, 1985.

MONTESI, M. *Método de evaluación y calidad de resúmenes documentales*. Gijón: Trea, 2006.

PINTO MOLINA, M. *El resumen documental*: principios y métodos. Madrid: Fundación Germán Sánchez Ruipérez, 2001.

REDIGOLO, F. M.; FUJITA, M. S. L. A leitura profissional do catalogador e seu papel como mediadora da informação. *Informação & Informação*, v. 20, n. 3, p. 356 - 376, set./dez. 2015.

RUBI, M. P. *A política de indexação na perspectiva do conhecimento organizacional*. Dissertação (Mestrado em Ciência da Informação) - Universidade Estadual Paulista, 2004.

RUBI, M. P. *Política de indexação para construção de catálogos coletivos em bibliotecas universitárias*. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista, 2008.

RUBI, M. P.; FUJITA, M. S. L.; BOCCATO, V. R. C. Elaboração do manual de política de indexação na formação continuada do catalogador. In: GIL LEIVA, I.; FUJITA, M. S. L. (Eds.). *Política de indexação*. Marília: Oficina Universitária, 2012. p. 217-227.

SILVEIRA, F. J. N.; MOURA, M. A. A estética da recepção e as práticas de leitura do bibliotecário-indexador. *Perspectivas em Ciência da Informação*, v. 12, n. 1, p. 125-135, jan.abr. 2007.

SMIT, J. W. (Coord.). *Análise documentária*: a análise da síntese. 2.ed. Brasília: IBICT, 1989.

SMIT, J.; BARRETO, A. A. Ciência da informação: base conceitual para a formação do profissional. In: VALENTIM, M. L. P. (Org.). *Formação do profissional da informação*. São Paulo: Polis, 2002. p. 9-23.

TABA, L. S.; CASELI, H. M. *Automatic semantic relation extraction from portuguese texts*, 2014. Disponivel em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/522_Paper.pdf>. Acesso em: 02 abr. 2017.

TABA, L. S. *Extração automática de relações semânticas a partir de textos escritos em português do Brasil*. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, 2013.