Uso das linguagens controlada e natural em bases de dados: revisão da literatura

Ilza Leite Lopes
Pesquisadora associada júnior na UnB, consultora em bancos de dados internacionais, professora na AEUDF, mestre em ciência da informação pela Universidade de Brasília.

Resumo

O trabalho tem como objetivo examinar o uso da linguagem controlada ou da linguagem natural , no planejamento da estratégia de busca em um ambiente de bases de dados em CD-ROM ou em linha. São revisados os estudos que abordam o uso das linguagens controlada e natural nas estratégias de busca, suas vantagens e desvantagens, proporcionando uma perspectiva sobre a complexidade para a busca da informação bibliográfica e referencial, incluindo a seleção de termos para as estratégias e a função do vocabulário controlado ou da linguagem natural nesse contexto.

Palavras-chave

Bases de dados; Estratégia de busca; Linguagem controlada; Linguagem natural. Recuperação da informação; Artigo de revisão.

Controlled and natural language use in the databases: literature review

Abstract

The work has the aim to examine the use of the controlled language or the use of the natural language in the search strategy planning in the CD-ROM or online environment. Studies that focusing the use of the controlled and natural languages in the search strategy are reviewed providing a perspective on the complexity of searching for bibliographic and referral information, including the selection of terms for the search strategies and the role of controlled vocabulary or the natural language in this context.

Keywords

Databases; Search strategy; Controlled language; Natural language; Information retrieval; Review article.

INTRODUÇÃO

A presente revisão não pretendeu realizar o mapeamento completo da extensa literatura que versou sobre os estudos de uso das linguagens natural e controlada como instrumentos de recuperação da informação. As vantagens e desvantagens das linguagens controlada e natural na busca de informação em bases de dados são apresentadas e, nesse contexto, o uso na estratégia de busca dos conceitos baseados nessas linguagens tem sido objeto de estudos com o objetivo de aprimorar os sistemas de recuperação da informação. Na recuperação da informação, utilizando-se bases de dados em cd-rom ou nos sistemas conversacionais, a linguagem controlada caracteriza-se como a que é utilizada apenas nos campos de descritor, termos de indexação e identificadores, sendo que a linguagem natural abrange os termos do título e do resumo dos documentos referenciados. Por esse motivo, a estratégia de busca precisa refletir a necessidade de informação do usuário.

A estratégia de busca pode ser definida como uma técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados. Isto significa que, a partir de um arquivo, um conjunto de itens que constituem a resposta de uma determinada pergunta será selecionado. Entretanto, a escolha de qual banco/base de dados é o mais adequado, a decisão sobre a linguagem a ser empregada, se a natural ou se a controlada, bem demonstra que o planejamento e execução de uma estratégia de busca envolve escolhas que irão determinar a obtenção eficaz daquelas informações específicas solicitadas pelo usuário ao serviço de informações. Para alcançar a resposta pretendida pelo usuário de informação, faz-se necessária a execução de movimentos e operações táticas, ora restringindo os resultados alcançados, ora ampliando-os para a obtenção de informações mais relevantes, conforme o pedido de busca demandado.

A literatura especializada registrou, desde os meados da década de 80, um forte movimento no sentido de prover o usuário final com os instrumentos necessários para que o mesmo realizasse suas próprias buscas. Interfaces amigáveis foram desenvolvidas para permitir o uso maciço dos bancos e bases de dados, bem como os programas baseados em "menus" onde o usuário menos experiente era orientado

passo a passo. Entretanto, os custos de conexão com as bases e com os bancos de dados e, no caso do Brasil, também os custos por caracteres transmitidos e recebidos inviabilizaram a participação do usuário final no processo de execução da estratégia de busca.

Apesar dos intensivos programas de treinamento oferecidos pelos produtores das bases de dados, pelos próprios sistemas de recuperação em linha, de toda a documentação existente sobre as características de cada base de dados e suas respectivas estruturas de informação, dos sistemas amigáveis que oferecem "menus" para guiar o usuário em cada etapa do processo de busca, das linguagens de busca com recursos especiais para se aproximarem cada vez mais do usuário inexperiente, o processo de busca continua sendo um fator de dificuldade que ainda não foi minimizado pelas novas tecnologias disponíveis.

ESTUDOS SOBRE AS LINGUAGENS: NATURAL E CONTROLADA

Ao longo das três últimas décadas de crescimento, expansão, utilização local e remota da informação armazenada em grandes sistemas de recuperação da informação, inúmeros autores se dedicaram ao estudo das linguagens natural e controlada aplicadas à indexação e recuperação de informações.

A linguagem natural (LN) pode ser definida como a linguagem do discurso técnico-científico, e, no contexto da recuperação da informação, Lancaster (1993, p. 200) afirma que "a expressão normalmente se refere às palavras que ocorrem em textos impressos, considerando-se como seu sinônimo a expressão "texto livre". Nas bases de dados, os campos de título e resumo registram os termos da LN, enquanto os campos de descritores, termos de indexação ou identificadores registram os termos da linguagem controlada (LC). Esta, denominada também vocabulário controlado, pode ser definida como um conjunto limitado de termos autorizados para uso na indexação e busca de documentos.

Bhattacharya (1974) analisou os experimentos realizados com a LN e o seu desempenho para recuperação nas áreas de aerodinâmica, ciência nuclear, física e biologia. Esses testes de avaliação do desempenho da LN tanto no processo de indexação, quanto no de busca, demonstraram que o uso da LN como linguagem de indexação e/ou de recuperação é viável com um controle mínimo de terminologia, ou mesmo com total ausência de controle nessas áreas. A autora demonstrou, ainda, que, nas áreas de química, física, botânica, zoologia e geologia, uma linguagem artificialmente elaborada, com controle de terminologia, torna-se quase impossível de ser criada devido ao acelerado desenvolvimento dessas áreas. Sugeriu, portanto, o uso da LN como instrumento de indexação e de recuperação simultaneamente. Citou também o esforço que, desde 1982, a International Union of Pure & Applied Chemistry (Iupac) vem desenvolvendo na área da química para o controle da nomenclatura dos compostos orgânicos, conforme extrato da introdução de publicação editada pela Iupac:

"Essas regras... constituem-se em recomendações para a denominação dos tipos de compostos e de compostos individuais. Elas não são exaustivas, excepto em casos específicos... A Comissão deseja que cada nação tente reduzir as variações de nomenclatura". (*apud* Battacharya, 1974, p.248).

Svenonious (1976), em um dos primeiros questionamentos sobre o processo de recuperação da informação, analisou os significados dos vocabulários controlados e da LN nas bases de dados em linha. Avaliou uma possível utilização dos conceitos que compõem um vocabulário controlado com o uso de medidas nas quais eles possam vir a ser quantificados e tratados como variáveis para serem utilizados em pesquisas teóricas e experimentais. Afirmou que o controle do vocabulário implica um processo classificatório, com duas etapas distintas. A primeira refere-se à classificação de variantes gramaticais do mesmo termo e/ou conceito, significado singular e plural, variantes gramaticais e diferentes flexões dos tempos verbais. Em uma segunda etapa, os termos e/ou conceitos são agrupados por descreverem o mesmo conceito ou um similar, isto é, sinônimos ou palavras que são equivalentes em seus significados.

Existem outras relações que podem ser úteis para o objetivo do controle do vocabulário, especialmente os relacionados com os processos de recuperação da informação. Assim, a possibilidade de truncagem de termos à direita permite a união automática de termos ortograficamente similares, podendo os mesmos serem ainda explorados em relação aos seus sufixos ou infixos. Os dicionários *on-line* oferecidos pelos melhores sistemas de bancos de dados apresentam em ordem alfabética termos com raízes idênticas e com suas variações gramaticais e respectivos significados diferentes, o que permite a seleção dos termos que podem ser aplicados na estratégica de busca. A autora finalizou afirmando que o controle bibliográfico não pode existir sem o controle do vocabulário, no que concordamos plenamente. Porém, essa questão, até o momento, ainda não foi solucionada, mesmo com as sofisticadas tecnologias de computação em vigor. Basta observar o uso dos robôs de busca utilizados atualmente na Internet que bem

demonstra o esforço de cada empresa em organizar e classificar o "caos" informacional circulante nas redes mundiais. Esses robôs criam e agrupam, em grandes classes, os milhões de sítios oferecidos na Internet, em uma tentativa de organização desse conhecimento.

Carrow & Nugent (1977) apresentaram uma avaliação comparativa dos métodos de busca com termos da LC *versus* termos da LN usando a base *National Criminal Justice Reference*. O texto pesquisado incluía resumos, títulos dos documentos e anotações. Os resultados mostraram que os dois métodos de busca apresentaram a mesma precisão no desempenho, mas as buscas com termos da LC produziram uma significativa e melhor recuperação. Os autores propuseram que os dois métodos fossem utilizados como complemento um do outro e afirmaram que o melhor desempenho da estratégia de busca seria aquele que utilizasse os dois métodos concomitantemente.

Henzler (1978) comparou amostras do uso de termos da LC e termos da LN na base de dados Cancernet. Em uma amostra de 100 títulos, ele encontrou que 35% de todas as palavras freqüentes nos títulos não tinham significados equivalentes, contra os descritores do vocabulário controlado. Confirmando que 50% dos descritores assinalados pela LC não possuíam representações na LN dos documentos, Henzler concluiu que tanto os termos da LN, quanto os do vocabulário controlado deviam estar presentes em uma combinação ideal durante a elaboração da estratégia de busca.

Um dos primeiros estudos da década de 80 sobre a utilização da LN e da LC na estratégia de busca foi o de Calkins (1980). Comparando as duas linguagens e utilizando as bases Compendex & Enviroline, a autora estabeleceu algumas hipóteses para a análise de um tema de interesse da U. S. Environmental Protection Agency (EPA). Na primeira hipótese, a premissa era que a busca em linha, em linguagem natural, recuperaria automaticamente todos os itens indexados (com a linguagem controlada) e, na segunda, que a busca utilizando o vocabulário controlado das bases mencionadas recuperaria virtualmente todos os itens mais pertinentes. Duas estratégias de busca foram elaboradas para a comprovação dessas hipóteses: uma baseada na LN e outra composta de termos de indexação da LC e códigos. Relatando as dificuldades encontradas para a escolha de termos, tanto para a LN quanto para a LC, a autora conclui que não pôde ser comprovada a primeira hipótese. O teste da segunda hipótese comprovou que as melhores citações foram recuperadas, porém em número bastante limitado. A conclusão foi pelo uso das duas linguagens simultaneamente na estratégia de busca, pois a combinação das mesmas aumentou consideravelmente a recuperação. Concluiu-se também que o intermediário "analista da busca" deve possuir grande habilidade em traduzir a necessidade de informação do usuário tanto para a linguagem de busca do sistema, quanto para as características de cada base.

Raitt (1980) apresentou a LC como a principal componente na estratégia de busca em sistemas *on-line* de recuperação de informação. Abordou os vários artifícios usados pelas linguagens controladas, para o aumento da revocação e da precisão. Dentre os que aumentam a revocação, nomeou os de controle de sinonímia, os de conexão de termos, controle de formas dos termos e os de agrupamento. Para os que influenciam a precisão, destacou os de coordenação, os de conexão e indicadores de função, bem como os que estabelecem pesos para os descritores. Revisou algumas linguagens controladas, dentre elas: *NASA Thesaurus*, *Thesaurus of Engineering and Scientific Terms*, *Thesaurus of Metallurgical Terms*, *Subject Headings for Engineering*, *INIS Thesaurus* e *INSPEC Thesaurus* .

Markey, Atherton & Newton (1980), no projeto de estudos sobre a base de dados ERIC *(Educational Resources Information Center*), no final da década de 70, analisaram o uso da LN e da LC nas estratégias de busca em linha. Nesse projeto, foram comparadas as buscas que utilizavam LN e a LC, objetivando comprovar, entre outras, as seguintes possibilidades: que características distinguem o vocabulário de busca? Que percentual da estratégia é formulado em LN? Qual a preferência pelo uso do termo da LN quando existe o termo da LC disponível? As análises relacionadas ao projeto não foram consideradas exaustivas pelas autoras, porém "pelo menos puderam proporcionar algumas indicações referentes a que tipos de tópicos de busca são melhores para uso do termo livre", já que um dos objetivos era aprimorar o acesso à base de dados ERIC, criando outros instrumentos de auxílio além do *thesaurus*.

Os resultados das estratégias de busca utilizando termos da LN e da LC foram comparados e submetidos a especialistas em educação. Essa análise revelou que as formulações de busca utilizando a LN tiveram maior revocação (93%) e menor precisão (71%) do que as estratégias que utilizaram a LC, com revocação de 76% e precisão de 95%. Concluíram que a busca com termos da LN pode, freqüentemente, ser a melhor opção quando se deseja alta revocação, todavia o uso combinado da LN e da LC oferece melhores resultados.

Analisando a crescente utilização da editoração eletrônica aplicada principalmente aos jornais de grande porte, Perez (1982) apresentou as vantagens e desvantagens do uso da

LN e da LC no registro de informações em bases de dados textuais. Este autor sugeriu que, através da geração de um pequeno vocabulário controlado, devia ser levado em consideração o enriquecimento das matérias jornalísticas, a fim de aumentar os benefícios da recuperação textual com o uso da indexação, já que a tendência é, cada vez mais, existirem bases de dados com texto completo.

Knapp (1982) observou que a estratégia de busca em LN é uma parte essencial no conhecimento dos intermediários que executam as buscas em bases de dados. Analisando as diferentes possibilidades de planejamento de estratégias de busca com a utilização da LN e com a LC, sugeriram algumas técnicas para elaboração das estratégias, focalizando principalmente as que usam a LN. Relacionaram-se, entre outras, os operadores de proximidade que podem ser usados nas buscas, os casos especiais para uso dos termos livres na estratégia, os problemas que ocorrem com a busca em LN, as fontes de pesquisa de sinônimos dos termos em LN e as estratégias para encontrar os termos nos bancos e nas bases de dados.

Schroder (1983) confirmou e acrescentou, na afirmação de Calkins (1980), que, além da LN e da LC, é preciso usar os acrônimos, quando for necessário. Alertou, porém, para a estrutura de informação de cada base. A identificação apropriada dos elementos descritivos de um item e/ou registro de informação contido em uma base de dados é de fundamental importância no planejamento da estratégia de busca. Vários autores têm se concentrado em estudos das estratégias de busca com o uso simultâneo da LN e da LC, como Dubois (1987) e Betts & Marrable (1991).

Wagers (1983) analisou a eficiência da estratégia de busca na LN por meio do desempenho unicamente dos termos que compõem o resumo dos documentos. Em experimentos realizados no Sistema Dialog e nas bases de dados *Energyline, Management e America History and Life,* o autor levantou a hipótese de que os termos livres das frases do resumo podem ser equiparados com os termos da LC e, nesse caso, podem ser usados eficientemente na estratégia de busca. Sugeriu ainda que, sob certas circunstâncias, o resumo pode concorrer mais eficientemente para busca em linha principalmente quando existe, por parte do produtor da base, interesse em destacar, por exemplo, instituições, dados factuais e outros.

Sievert & Boyce (1983) levantaram um questionamento sobre a função tradicional da LC como instrumento de recuperação. Analisando a estrutura de informação das bases e do banco de dados Dialog, demonstraram que as linguagens controladas estão sendo efetivamente utilizadas como um instrumento de precisão, e não de revocação. Nos testes levados a efeito com termos extraídos do *Thesaurus of ERIC Descriptors*, da base ERIC, já analisada por Markey, Atherton & Newton (1980), os autores concluíram que a LC pode ser eficientemente utilizada sem a completa entrada dos termos, tendo em vista a possibilidade de combiná-los com os operadores de adjacência do sistema Dialog, e com os recursos do sistema para restrição do uso desse termos, apenas no campo de descritores.

Dentre as linguagens controladas que podem ser aplicadas na recuperação da informação, Svenonious (1983) destacou, como um instrumento de busca, o uso do sistema de classificação definido pelo autor da base. Uma das possíveis formas em que a classificação pode ser usada na recuperação é a ampliação do número de documentos relevantes recuperados. Essa classificação, segundo a autora, não se restringe às tabelas de classificação tradicionais, antes inclui todos os instrumentos classificatórios desenvolvidos e aplicados pelos produtores das bases de dados no processo de indexação. Um bom exemplo desse instrumento é a base correspondente ao *Biological Abstracts* impresso, denominada *Biosis Previews*, que utiliza uma lista de palavras-chave, uma de códigos de conceitos e outra de códigos biossistemáticos para a indexação ou recuperação de informação.

Piternick (1984) destacou alguns tipos de vocabulários utilizados tradicionalmente como instrumentos de indexação que, com o advento dos sistemas de recuperação *on-line*, foram transformados em vocabulários de busca. Afirmou que os "*Thesauri* e Listas de Cabeçalhos de Assunto, em um passado recente, caracterizavam-se mais como vocabulários de indexação do que como vocabulários de busca", propondo, então, a geração de vocabulários voltados para a recuperação da informação, e não para a indexação. Alertou, ainda, que, embora um vocabulário controlado seja extremamente útil para elaborar estratégias de busca visando especificamente à precisão dos resultados, existem momentos em que se torna imprescindível a utilização de termos extraídos da LN.

Austin (1986) relatou as mudanças registradas nas funções específicas de indexação com o uso dos computadores, mostrando que as mudanças ocorreram também com novas atividades na geração dos produtos de indexação: os índices e os resumos. Relembrou que o uso de vocabulário controlado na indexação e o seu respectivo uso na recuperação da informação vão requerer o estabelecimento de certas regras terminológicas recomendadas, tais como:

"Conceitos devem ser representados consistentemente para os propósitos de recuperação, por substantivos ou frases substantivadas; os indexadores devem trabalhar com um vocabulário de termos preferidos, designando-se um dos sinônimos de um determinado conceito como o termo mais adequado para uso; a opção pelo singular ou plural dos conceitos e suas exceções devem ser registradas claramente nesse vocabulário, visando à consistência da indexação, sua fidedignidade e posterior uso na recuperação". (Austin, 1986, p.8)

Essa preocupação de Austin, no auge do crescimento da indústria da informação, com a geração dos índices e resumos em forma legível por computador, denominadas bases de dados e sua conseqüente comercialização com o acesso remoto e, também, com o mercado crescente do CD-ROM, denota o necessário desenvolvimento de instrumentos de apoio para a elaboração das estratégias de busca, que são os vocabulários controlados ou as linguagens documentárias geradas pelo produtor de cada base de dados.

Harter (1986) analisou as linguagens para recuperação da informação, abordando suas características e classificando-as como linguagens de busca, linguagem natural, indexação por assunto, vocabulários controlados e indexação por citações. Segundo o autor, tanto o uso da LC quanto o uso da LN apresentam vantagens e desvantagens na indexação e na recuperação da informação. Enquanto o primeiro é "rígido, inflexível, mas preciso, o outro é altamente expressivo, flexível, mas potencialmente ambíguo.

Fidel (1986) analisou a importância dos guias para elaboração de resumos desenvolvidos pelos produtores de bases de dados, com vistas ao aprimoramento de busca de informações em linguagem natural. Comprovou que o desempenho na recuperação do documento está diretamente relacionado com o conteúdo e o estilo da linguagem utilizada no resumo. Em outro estudo, também de 1986, a autora investigou as políticas de resumo usadas pelos produtores das bases de dados com o objetivo de identificar a relação entre a ampliação e a recuperação de informação na LN contida nos resumos dos documentos. Finalizou, afirmando que um guia para a elaboração de resumos com seus diversos tipos e determinando ainda os respectivos tamanhos dos mesmos pode oferecer melhor desempenho na recuperação com a LN.

Tenopir (1987) examinou a postura do intermediário que operacionaliza as buscas e reafirmou que, independentemente da experiência adquirida, o mesmo se posiciona de forma estática no processo de recuperação, pois age como um profissional que não questiona o processo subjetivo de indexação, selecionando para a estratégia de busca apenas termos da LC. Mencionou que muitas bases de dados não desenvolveram linguagens controladas para a recuperação, porém as que são subproduto de acervos de grandes bibliotecas quase sempre já geraram, entre outros, os cabeçalhos de assunto, *thesaurus* ou ainda listas de descritores permitidos para a indexação e/ou recuperação. A autora apontou várias dificuldades a serem superadas em relação ao uso apenas da LC nas buscas, causadas em grande parte pela atualização constante do conhecimento e da morosidade em se manterem atualizados os respectivos vocabulários. Sugeriu que, dependendo do tema de busca solicitado e da base de dados a ser consultada, sejam utilizadas nas mesmas estratégias tanto nos termos da LN, quanto nos termos da LC.

Svenonious (1988), em um contexto de recuperação de informação, analisou os instrumentos utilizados para indexação (a saber: as classificações, os cabeçalhos de assuntos e os *thesaurus*), apresentando considerações para a construção dos mesmos. Esses vocabulários controlados e suas respectivas funções nos sistemas de recuperação da informação contribuem para a ampliação dos resultados relevantes desejados pelos usuários. Recomendou o desenvolvimento de vocabulários controlados, voltados para a recuperação da informação, e relembrou que bases de dados podem ser agrupadas em grandes bases, porém torna-se necessário compatibilizar as regras para a geração de *thesaurus*, visando à normalização do conjunto de vocabulários controlados.

Lancaster, Elliker & Connel (1989), em um artigo de revisão do Arist no período de 1986-1988, agruparam os documentos analisados em seis grandes categorias: teoria e prática da indexação; vocabulários controlados, incluindo classificação e cabeçalhos de assunto; estratégias de busca e métodos de busca; busca em linguagem natural; indexação automática e o uso de citações na recuperação da informação.

Boyce & McLain (1989), em um estudo realizado na linguagem de recuperação de bancos de dados, revelaram que a profundidade da indexação, incluindo o total de termos que são assinalados em média por documento, e os pontos de acesso disponíveis no banco têm expressivo efeito nos padrões de desempenho desses sistemas de recuperação da informação. Demonstraram que, como os termos da LC aumentam a precisão nos resultados da busca, o seu uso em uma estratégia deve ser criteriosamente examinado, tendo em vista a estrutura de informação oferecida pelo sistema.

Nicholls & Holtmann (1989) relataram a investigação realizada em uma base de dados em CD-ROM na qual foram testados o uso da LN e da LC na estratégia de busca sobre um tema específico, demonstrando que os dois enfoques ainda são muito polêmicos no que se refere aos fatores de revocação e precisão. Sugeriram, para a estratégia de busca, uma combinação flexível de termos extraídos da LN e da LC da base de dados a ser utilizada para a pesquisa, lembrando que a natureza do pedido de busca, a estrutura de informação de base e a linguagem de busca do sistema são variáveis intervenientes no processo de recuperação.

Comber & Stanford (1989) estabeleceram analogias entre o conceito de texto-livre, ou LN, e o conceito de LC, em um contexto de busca em linha. Mencionaram as dificuldades inerentes ao processo de produção de bases de dados no que se refere às diretrizes a serem adotadas nas atividades de indexação, entre elas, os fatores custo de indexação e de recuperação.

Em um projeto desenvolvido em 1987, Rowley (1990) comparou a indexação e busca na LN com a busca na LC. Utilizando uma pequena base de dados contendo títulos e descritores extraídos do *ERIC Thesaurus* e os cabeçalhos de assunto da *Sears List of Subject Headings*, a autora realizou uma série de buscas usando termos da LN e da LC. As medidas de revocação e precisão foram comparadas no experimento para avaliar a eficácia de cada uma das linguagens durante a busca. Comprovou que o uso de uma linguagem de indexação controlada na busca não ofereceu, necessariamente, o melhor desempenho. Afirmou que a disponibilidade de novas interfaces para o uso da LN na busca de informações com a introdução dos sistemas hipermídia e hipertexto poderá provocar um desempenho melhor tanto na LN quanto na LC para a busca da informação.

Fidel (1991) relatou o experimento realizado com os intermediários durante o processo de seleção da terminologia a ser utilizada na estratégia denominada "chaves-de-busca". Analisou o uso de descritores extraídos de vocabulários controlados e também a busca que utiliza texto livre.

Kaback (1992) analisou um dos debates mais controversos no contexto de recuperação da informação que tem sido o argumento sobre o uso ou não da LN ou da LC nas estratégia de busca. Nessa investigação em bases de patentes foram testadas as hipóteses sobre o uso da LN e da LC, é, para cada base consultada sobre o tema, foram levantadas as dificuldades relativas aos tipos de vocabulários possíveis de serem utilizados na estratégia de busca. Concluiu que, para a área de patentes, devem ser usadas ambas as linguagens na estratégia de busca, incluindo-se, ainda, a Classificação Internacional de Patentes.

Rowley (1994) revisou o debate sobre a LN e a LC utilizadas no contexto de indexação e recuperação das bases de dados em linha dos sistemas de recuperação da informação, dos catálogos em linha de acesso público e das bases de dados em CD-ROM. A prática e a experiência profissional têm demonstrado o quão distante os intermediários ou os usuários finais se encontram desses instrumentos de trabalho, principalmente no que se refere à recuperação de informação. Os próprios bancos de dados e os produtores das bases de dados nem sempre apresentam esses documentos ao público usuário dos seus serviços, o que dificulta sobremaneira o planejamento da estratégia de busca. Existem exceções, como o banco de dados DIALOG e o ORBIT, que incluem no manual do banco de dados um capítulo denominado *"Search Aids"*, no qual relacionam os diversos vocabulários controlados utilizados na indexação dos documentos.

LC E LN: VANTAGENS E DESVANTAGENS

Em princípio, a forma exata de uma estratégia de busca é determinada pela natureza da base de dados a ser acessada e pela sua respectiva estrutura de informação, isto é, pela formatação de seus campos de identificação do documento e dos campos de identificação do conteúdo temático do mesmo. A identificação dessa estrutura de campos de busca implica o conhecimento da documentação básica fornecida pelos produtores das bases e pelos bancos de dados onde estão hospedadas. Nas bases em CD-ROM, a maioria dos produtores armazena nos próprios discos a LC das mesmas, facilitando o planejamento da estratégia de busca.

Lancaster (1979) observou que, embora os computadores possibilitem a manipulação de extensas listas de palavras, ainda não contribuem para a solução dos problemas intelectuais relacionados com a elaboração das estratégias de busca em linguagem natural, pois:

"O thesaurus, ou outro instrumento de controle do vocabulário oferece muita ajuda aos intermediários que executam a busca, incluindo, entre outros, o controle de sinônimos e quase-sinônimos; a separação dos homógrafos; o uso da pré-coordenação para evitar falsas coordenações e relações incorretas entre os termos e, ainda, a ligação de termos relacionados , todos com suas respectivas hierarquias.... O thesaurus pode ser muito específico, mas nunca tão específico quanto a linguagem natural, que é a linguagem do discurso dos próprios autores." (Lancaster, 1979, p. 281)

Rothman (1983) levantou um questionamento pertinente sobre a busca de informação na LN, em oposição à busca com a LC. Afirmou que, com a busca na LN, a base de dados está efetivamente auto-indexada, pois cada palavra no documento indexado é candidata a termo de busca e identifica, ainda, a unidade do texto no qual se encontra. Por esse motivo, a LN dos documentos constitui termos de indexação ou pontos de acesso imediato, e os usuários podem interagir diretamente com os itens incluídos base, enquanto, com a busca na LC, o indexador é interposto entre os usuários e a base de dados, ficando na posição de mediador ou intérprete.

Naturalmente que o uso da LN tem dificuldades a serem superadas, da mesma forma que a LC apresenta certas desvantagens, pois os termos preferidos pelos indexadores freqüentemente não são os termos utilizados pelos usuários em situações específicas de busca, pois, em grandes bases de dados, o processo de indexação envolvendo diversos indexadores provoca, com certeza, inconsistências na identificação de documentos similares. Essas discrepâncias entre termos assinalados pelos indexadores e os termos utilizados pelos usuários no momento de busca não podem ser considerados genericamente erros, porque, na realidade, o processo de indexação ocorre em outro contexto, o de análise de conteúdo do documento e a tradução desse conteúdo para o vocabulário controlado da base. A busca de informação do usuário, além disso, precisa ser traduzida para a linguagem controlada da base, pelo próprio usuário ou por um intermediário. Portanto, as deficiências são inevitáveis, principalmente quando ambos desconhecem as linguagens controladas da(s) base(s) de dados a serem consultadas.

LINGUAGEM CONTROLADA OU VOCABULÁRIO CONTROLADO

Pode ser definido como um conjunto de termos organizados de forma hierarquizada e/ou alfabética, com o objetivo de possibilitar a recuperação de informações temáticas, reduzindo substancialmente a diversidade de terminologia. São também conhecidos como linguagens documentárias ou linguagens controladas. Uma base de dados que utilize um vocabulário controlado possibilita, ao intermediário no planejamento da estratégia de busca, a recuperação, no campo específico de descritor, apenas daquelas palavras-chave listadas no *thesaurus* e/ou vocabulário controlado da base de dados.

TABELA 1
Vocabulário controlado: vantagens e desvantagens

VANTAGENS	DESVANTAGENS
1. Controle total do vocabulário de indexação, minimizando os problemas de comunicação entre indexadores e usuários.	Custos: a produção e manutenção da base de dados terá despesas maiores com a equipe de indexadores. Será necessário ainda manter pessoal especializado na atualização do *thesaurus*.
2. Com o uso de um *thesaurus* e suas respectivas notas de escopo, os indexadores podem assinalar mais corretamente os conceitos dos documentos.	O vocabulário controlado poderá não refletir adequadamente os objetivos do produtor da base, caso esteja desatualizado.
3. Se bem constituído, o vocabulário controlado poderá oferecer alta recuperação e relevância e, também, ampliar a confiança do usuário diante de um possível resultado negativo.	Um vocabulário controlado poderá se distanciar dos conceitos adequados para a representação das necessidades de informação dos usuários.
4. As relações hierárquicas e as remissivas do vocabulário controlado auxiliam tanto o indexador, quanto o usuário na identificação de conceitos relacionados.	Necessidade de treinamento no uso dos vocabulários controlados tanto para os intermediários, quanto para os usuários finais.
5. Redução no tempo de consulta à base, pois a estratégia de busca será mais bem elaborada com o uso do *thesaurus*.	Desatualização do vocabulário controlado poderá conduzir a falsos resultados.

Revendo o processo de busca de informação usando o vocabulário controlado, Brundage (1989) traçou um paralelo entre a linguagem científica e a linguagem controlada. Relembrou que os cientistas de uma certa forma estão familiarizados com o vocabulário controlado em suas áreas de especialização, portanto os paradigmas dos conhecimentos recebidos nas disciplinas podem ser usados como modelo no ensino do uso, especialmente, dos *thesauri*. Observou que o intermediário que operacionaliza as buscas dispõe de conhecimentos prévios sobre o uso da LC e da LN, porém o usuário final possui uma profunda intuição sobre a terminologia de seu campo de especialização, e essas diferenças vão orientar os programas de treinamento para os diversificados grupos de participantes. As principais vantagens e desvantagens do uso do vocabulário controlado na recuperação da informação mencionadas por Henzler (1978), Perez (1982), Salton (1986) e outros são sintetizadas na tabela 1.

LINGUAGEM NATURAL OU VOCABULÁRIO LIVRE

Conceitua-se a expressão linguagem natural como sinônimo de discurso comum, isto é, a linguagem usada habitualmente na fala e na escrita sendo que, nas bases de dados, os termos do título e resumo representam a LN.

Knapp (1982) resumiu brevemente os casos em que a busca em LN pode obter melhor desempenho: para tópicos específicos; para temas atuais; para novas terminologias ainda não incluídas nas LC; para uma busca retrospectiva em que o conceito da LC é muito recente e não cobre os anos anteriores; quando o termo da LC é muito abrangente ou muito específico; para pesquisa em várias bases de dados; para identificação imediata de palavras de títulos dos documentos; para complementação de citação bibliográfica incompleta.

Nas bases de dados bibliográficas, os campos de busca em que se pode pesquisar usando apenas termos e/ou conceitos da LN normalmente são os do título e resumo dos documentos. Nesses campos, cada palavra é automaticamente candidata a ser pesquisada, excetuando-se aquelas designadas pelos sistemas como não-significativas, as quais vão compor as listas de palavras proibidas. Assim, as palavras remanescentes são usadas para criar índices que podem ser pesquisados na LN. Esse recurso pode ser utilizado para rastrear temas e seus respectivos conceitos terminológicos que ainda não foram incluídos na LC, ou seja, nos *thesauri*, nas listas de cabeçalhos de assunto, nas tabelas de classificações especializadas, listas de descritores, códigos taxonômicos, nomenclaturas e outros. Cabe ao intermediário que operacionaliza as estratégias de busca a decisão tática de utilizar esses recursos no planejamento das estratégias, para alternativamente ampliar ou restringir os resultados que se pretendem, de acordo com o definido pelo usuário da informação. As principais vantagens e desvantagens da LN citadas na literatura especializada mencionada anteriormente são apresentadas na tabela 2.

TABELA 2
Linguagem natural: vantagens e desvantagens

VANTAGENS	DESVANTAGENS
1. Permite o imediato registro da informação em uma base de dados, sem necessidade de consulta a uma linguagem de controle.	Os usuários da informação, no processo de busca, precisam fazer um esforço intelectual maior para identificar os sinônimos, as grafias alternativas, os homônimos etc.
2. Processo de busca é facilitado com a ausência de treinamentos específicos no uso de uma linguagem de controle.	Haverá alta incidência de respostas negativas ou de relações incorretas entre os termos usados na busca (por ausência de padronização).
3. Termos de entrada de dados são extraídos diretamente dos documentos que vão constituir a base de dados.	Custos de acesso tendem a aumentar com a entrada de termos de busca aleatórios.
4. Temas específicos citados nos documentos podem ser encontrados.	Uma estratégia de busca que arrole todos os principais conceitos e seus sinônimos deve ser elaborada para cada base de dados (ex: nomes comerciais de substâncias químicas não ocorrem no Chemical Abstracts).
5. Elimina os conflitos de comunicação entre os indexadores e os usuários, pois ambos terão acesso aos mesmos termos.	Perda de confiança do usuário em uma possível resposta negativa.

LC E LN EM BASES DE DADOS

Os diversos sistemas de recuperação da informação denominados bancos de dados também participam do processo de indexação das bases de dados hospedadas nos mesmos. Assim, os bancos como DIALOG, ORBIT, BRS e outros preparam índices em linha para cada base de dados disponibilizada para consulta. Os produtores de bases de dados em cd-rom também participam do processo de indexação das bases fornecidas com suas interfaces de busca, oferecendo índices constituídos pelos próprios registros das bases de dados.

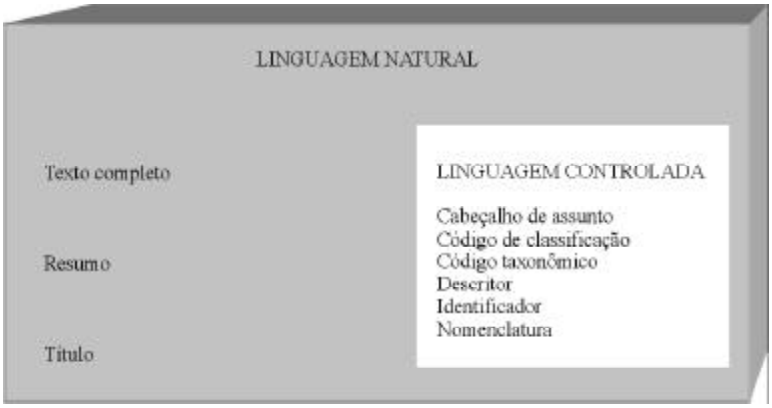
Anderson & Wilson (1983) confirmaram que os índices assim constituídos freqüentemente oferecem as seguintes possibilidades: habilidade para busca de informação em um campo específico; habilidade para busca em todos os campos de assuntos simultaneamente; habilidade para evitar referências de campos não desejados; políticas de análise gramatical para cada campo visando à maximização da capacidade de busca; consistência no tratamento de cada campo, incluindo apresentação dos registros em formatos predefinidos; facilidades para busca precisa de subcampos; consistência no tratamento de campos similares e habilidade para pesquisar simultaneamente o equivalente a múltiplos volumes impressos de diversos anos anteriores.

Knapp, Cohen & Juedes (1998) salientaram que a pesquisa em bases de dados apresenta dificuldades especiais para a área de humanidades, porque os assuntos podem ser abordados de inúmeras maneiras, sendo que vários sinônimos podem ser usados para descrever um só conceito e os termos podem apresentar variações relativas à precisão. Apontaram, como causa significativa na baixa recuperação de informação, a inabilidade dos intermediários e/ou dos usuários finais em selecionar todos os possíveis termos que os autores utilizaram nos seus trabalhos, o que já tinha sido observado por Lancaster & Fayen (1973). Apresentaram como proposta os resultados de um estudo utilizando estratégias de busca em LC e LN, no qual ficou demonstrado que a combinação dos dois modelos de estratégia usados em conjunto ofereceu maior recuperação. Em um estudo realizado na área de mecânica dos solos, Muddamalle (1998) demonstrou que o uso concomitante da LC e da LN nas estratégias de busca apresentou melhor desempenho na recuperação do que as estratégias que foram realizadas com cada tipo de linguagem individualmente. Recomendou que, a fim de serem obtidos resultados satisfatórios na recuperação, uma combinação na estratégia de termos da LC e da LN deve ser adotada. Salientou que:

"A LC e a LN não podem mais ser tratadas como técnicas de busca separadas, mas devem sempre ser tratadas em conjunto, como uma combinação ideal para ampliar os resultados das buscas de informação." (Muddamalle, 1998, p. 887)

Com o objetivo de identificar as similaridades semânticas no processo de recuperação da informação, Qin (2000) escolheu um assunto específico na área de saúde: resistência a antibióticos. Para essa investigação, foram definidas duas bases de dados: uma com LC e outra em LN. A que utiliza apenas a LN, enriquecida com palavras dos títulos das citações referenciadas nos artigos, foi a *Science Citation Index*. A base selecionada com a LC foi a Medline, que indexa seus documentos utilizando o *Medical Subject Headings (MESH)*. Os primeiros resultados do teste da estratégia com a LC e a LN sobre o tema escolhido demonstraram as limitações da indexação da base em LC, além de comprovar quão ampliada pode ser a dissimilaridade entre os diferentes métodos de indexação do mesmo documento.

FIGURA 1
Estrutura da informação para representação do assunto



Os termos que constituem o vocabulário de uma LC são de dois tipos: termos derivados e termos assinalados. Os primeiros são extraídos dos próprios documentos e os últimos são os termos utilizados para normalização bibliográfica, também denominados metadados de organização da informação. Os termos derivados descrevem a própria linguagem do autor, isto é, termos da LN, e os assinalados são os termos que constituem os vocabulários controlados estabelecendo as relações hierárquicas entre os componentes da LC. Na LC das bases, eles são denominados palavras-chave, descritores, termos de indexação, códigos de classificação e outros.

Svenonious (2000) afirmou que a LC é derivada da literatura que pretende descrever, portanto torna-se necessária a delimitação de suas fronteiras, tendo definido a LC como:

"Uma linguagem normalizada que seleciona da linguagem natural um conjunto restrito de palavras e frases, tratando esse conjunto semanticamente para o estabelecimento das relações com outros termos." (Svenonious, 2000, p. 134)

Salientou, ainda, que existem dois tipos principais de LC: a linguagem de assunto alfabética e a linguagem de assunto classificatória, apontando três diferenças entre elas. A primeira utiliza expressões verbais, *thesauri* e listas autorizadas de termos e ordena os assuntos alfabeticamente. A segunda usa expressões verbais e notações, esquemas de classificação e ordena assuntos sistematicamente, primeiro por disciplina e dentro desta, hierarquicamente, por tópicos.

A estrutura da informação para a representação do conteúdo temático dos documentos nas bases de dados em linha, nos respectivos bancos e nos produtos em CD-ROM, permite a consulta tanto de termos da LN, quanto de termos da LC, conforme apresentado na figura 1.

Em bases de dados, a representação do conteúdo dos documentos pode estar contida nos campos de busca da LN e da LC, portanto, considerando que esta é um subconjunto daquela, conforme sintetizado na figura 1, a visualização desses campos permite inferir que, em determinados contextos, os termos da LN e da LC podem ser superpostos.

Vários sistemas de recuperação de informação e produtores de bases em CD-ROM oferecem o acesso a múltiplas bases de dados, e, para cada área do conhecimento humano, várias bases estão disponibilizadas no contexto da indústria da informação. Entretanto, cada base utiliza o seu próprio vocabulário controlado. Uma vez que não existe padronização ou compatibilidade entre esses vocabulários, mesmo naqueles que são específicos de um determinado assunto, o mesmo conceito pode ser designado por diferentes termos nos distintos vocabulários. Svenonious (1986) confirmou essa falta de padronização e alertou para a função do *thesaurus* como instrumento imprescindível na recuperação da informação.

Duas técnicas especialmente úteis para a busca em LN são a de truncagem de termos e a busca com operadores de proximidade. A truncagem de termos permite ao intermediário que operacionaliza a busca usar a raiz do termo sem especificar todas as possíveis variações desse termo (prefixos e/ou sufixos). Já a técnica de busca com operadores de proximidade ou de adjacência permite especificar, na estratégia, a posição relativa de dois ou mais termos entre eles próprios. Nas buscas relativas ao levantamento das últimas tecnologias em uma determinada área, ou a um novo assunto, ou a um novo produto e, ainda, na busca em documentos de patentes, o uso da estratégia em LN pode ser o melhor caminho para o encontro da informação desejada. A terminologia pode ser muito atual, as aplicações ainda não são tão significativas para serem indexadas, portanto os novos termos não foram incorporados a nenhuma lista autorizada de LC.

Segundo Lancaster (1986), a LN "não tem atraso, nem vocabulário específico, mas tem a garantia bibliográfica total", pois os termos são extraídos diretamente dos documentos.

Harter (1986), comparando os problemas de ambigüidade da LN em áreas especializadas, como física e química, e em áreas de ciências sociais, como educação, ou em bases de dados multidisciplinares, confirmou o valor da LN para a pesquisa nas ciências exatas e alertou para as dificuldades a serem superadas nas áreas sociais e humanas. Complementando essas análises, o autor ressaltou que, na indexação com LC o número de conceitos indexados é relativamente menor, quando se comparam com os pontos de acesso permitidos pela LN, e que:

"Tipicamente, apenas alguns poucos termos são assinalados para representar o conteúdo dos documentos. Entretanto, mesmo a indexação exaustiva como a usada pela *National Library of Medicin*e não supera a exaustividade proporcionada pela linguagem natural, especialmente aquelas oferecidas pelo texto completo dos documentos." (Harter, 1986, p.54)

As discussões e estudos sobre o uso da LN e da LC na recuperação da informação vêm se estendendo há bastante tempo, e a solução ainda não foi encontrada. O desenvolvimento de linguagens controladas para indexação e recuperação da informação requer trabalho em equipe, altos investimentos e tempo para gerar, testar e sedimentar um instrumento de trabalho que deve ser continuamente atualizado. Com o foco da busca centrado no usuário final das bases em CD-ROM e dos OPACs, reconhece-se que ele tem excessiva dificuldade em formular estratégias de busca apropriadas e, portanto, não se beneficia do uso das linguagens controladas, ainda que as mesmas estejam documentadas e associadas com as bases de dados.

Harter (1986), Rowley (1994), Lancaster (1993) e outros estudiosos do assunto sugerem que a prática deve nortear a decisão da escolha dos termos da LN e da LC. Muitos intermediários e/ou usuários finais empregam ambas as linguagens no momento de formulação da estratégia de busca, principalmente porque, na maioria das bases de dados, é possível a busca simultânea por campos de busca que podem ser combinados entre si. Assim, os campos de resumo, de títulos, de identificadores, de descritores ou cabeçalhos de assunto e de códigos de classificação podem ser amplamente utilizados visando à obtenção de um resultado mais satisfatório, independentemente da verificação, no momento de operação da busca, de qual dessas linguagens terá melhor desempenho. O foco, portanto, está na obtenção de resultados satisfatórios, e não no instrumento utilizado para alcançar esses resultados.

CONCLUSÕES

Inúmeros estudos sobre o uso da LC e da LN na recuperação da informação têm se concentrado na utilização conjunta das duas linguagens na estratégia de busca, comprovando que o uso simultâneo dessas linguagens proporciona melhor desempenho nos resultados. Observou-se, também, a predominância de investigações que descrevem as vantagens e desvantagens do uso da LN e da LC nas estratégias de busca em bases de dados. Essa observação reflete a ampla difusão do tema, sem, contudo, apontar especificamente a melhor linguagem, pois vários outros fatores intervenientes afetam os resultados finais das buscas em bases de dados.

As dificuldades comuns, identificadas por Rowley (1990, 1994), Svenonious (1976, 1986, 2000), Harter (1986), Lancaster (1979, 1986, 1993) e Fidel (1987, 1991), entre outros, relacionadas com as regras de seleção dos termos para a estratégia de busca, indicam que a prática deve nortear a decisão final. Porém, o debate sobre o uso da LC ou da LN permanece sendo questionado até os dias atuais. Considera-se que a complexidade do problema deve proporcionar novos estudos que venham a contribuir para o aprimoramento do processo de decisão relativo à seleção de termos que vão compor as estratégias de busca em bases de dados, seja em bases de dados em CD-ROM, seja em bases de dados em linha.

Artigo aceito para publicação em 06-01-2002

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, J. D.; WILSON, C. B. Essential decisions in indexing systems design. In: FEINBERG, H. Indexing specialized formats and subjects. London : Scarecrow, 1983. cap. 1

AUSTIN, D. Vocabulary control and information technology. Aslib Proceedings, v. 38, n. 1, p. 1-15, Jan. 1986.

BATES, Marcia J. How to use controlled vocabulaires more effectively in online searching. Online, v. 12, n. 6, p. 45-56, Nov. 1988.

BATTACHARYA, K. The effectiveness of natural language in science indexing and retrieval. Journal of Documentation, v. 30, n. 3, p. 235-293, Sept. 1974.

BETTS,R.; MARRABLE, D. Free text vs controlled vocabulary – retrieval precision and recall over large databases.In: INTERNATIONAL ONLINE INFORMATION MEETING, 1991, London. Proceedings... London : Learned Information, 1991. p. 153-165.

BOYCE, B. R. , McLAIN, J. P. Entry point depth and online search using a controlled vocabulary. JASIS, v. 40, n. 4, p. 273-276, 1989.

BRUNDAGE, Christina A. Teaching controlled vocabulary and natural language to end-users. Science & Technology Libraries, v. 10, n. 1, p. 3-13, Fall 1989.

CALKINS, Mary L. Free text or controlled vocabulary? Online, v. 3, n. 2, p. 53-65, June, 1980.

CARROW, D.; NUGENT, J. Comparison of free-text and index search abilities in an operating information system. In: INFORMATION MANAGEMENT IN THE 1980's. 1977, New York. Proceedings... New Yorkv : ASIS, 1977. v. 14, p. 232-238.

COMBER, D.; STANFORD, J. A. Comparison between free text and a thesaurus controlled vocabulary. Records Management Journal, v. 1, n. 3, p. 113-120, Aug. 1989.

DUBOIS, C. P. R. Free text vs. controlled vocabulary. Online Review, v. 11, n. 4, p. 243-251, l987.

FENICHEL, C. H. The process of searching online bibliographic databases, a review of research. Library Research, v. 2, p. 107-127, 1980.

FIDEL, Raya. The possible effect of abstracting guidelines on retrieval performance of free-text searching. Information Processing and Management, v. 22, n. 4, p. 309-316, 1986.

_____. Searchers selection of search keys: I. The selection routine. JASIS, v. 42, n. 7, p. 490-500, 1991.

_____. Searchers selection of search keys: II. Controlled vocabulary or free-text searching. JASIS, v. 42, n. 7, p. 501-514, 1991.

_____. Searchers selection of search keys. III. Searching styles. JASIS, v. 42, n. 7, p. 515- 527, 1991.

_____. Writing abstracts for free-text searching. Journal of Documentation, v. 42, n. 1, p. 11-21, Mar. 1986.

HARTER, S. P. Online Information retrieval: concepts, principles and techniques. London : Academic Press, 1986. p. 22-63.

HENZLER, R. G. Free or controlled vocabularies. International Classification, v. 5, n. 1, p. 21-26, 32, Mar. 1978.

KABACK, S. M. Online patent information: who needs indexing? We do, naturally. World Patent Information, v. 14, n. 3, p. 198-199, Aug. 1992.

KNAPP, Sara D. Free-text searching of online databases. Reference Librarian, n. 5/6, p. 143-153, Fall/Winter 1982.

_____; COHEN, L. B.; JUEDES, D. R. A natural language thesaurus for the humanities: the need for a database search aid. Library Quarterly, v. 68, n. 4, p. 406-30, Oct. 1998.

LANCASTER, F. W. Indexação e resumos: teoria e prática. Brasília : Briquet de Lemos, 1993.

_____. Information retrieval systems: characteristics, testing and evaluation. 2nd ed. New York : Wiley, 1979.

_____. Vocabulary control for information retrieval. 2nd ed. Arlington : IRP, 1986.

_____; FAYEN, E. G. Information retrieval on-line. Los angeles : Melville, 1973.

_____; ELLIKER, C.; CONNEL, T. Subject analysis. ARIST, v. 24, p. 35-93, 1989.

MARKEY, K.; ATHERTON, P.; NEWTON, C. An analysis of controlled vocabulary and free text search statements in online searches. Online Review, v. 4, n. 3, p. 225-323, Sept. l980.

MUDDAMALLE, Manikya Rao. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. JASIS, v. 49, n. 10, p. 881-887, Oct. 1998.

NICHOLLS, P.; HOLTMANN, S. Women's issues searching with Dialog Eric: natural language and controlled vocabulary strategies. Laserdisk Professional, v. 2, n. 3, p. 97-103, 1989.

PEREZ, Ernest. Text enhancement: controlled vocabularies vs free text. Special Libraries, v. 72, n. 3/4, p. 183-192, 1982.

PITERNICK, Anne B. Searching vocabulaires: a developing category of online search tools. Online Review, v .8, n. 5, p. 441-453, Oct. l984.

QIN, Jian. Semantic similarities between a keyword database and a controlled vocabulary database: an investigation. JASIS, v. 51, n. 3, p. 166-180, Mar. 2000.

RAITT, D. J. Recall and precision devices in interactive bibliographic search and retrieval systems. Aslib Proceedings, v. 32, n. 7/8, p. 281-301, July/Aug. 1980.

ROTHMAN, J. Is indexing obsolete? Keyword indexing and free-text searching.In: FEINBERG, H. Indexing specialized formats and subjects. London : Scarecrow, 1983. cap. 2, p. 22-34.

ROWLEY, J. E. A comparison between free language and controlled language indexing and searching. Information Services & Use, v. 10, n. 3, p. 147-155, 1990.

_____. The controlled versus natural indexing languages debate revisited. Journal of Information Science, v. 20, n. 2, p. 108-119, 1994.

SCHRODER, J. J. Study of strategies used in online searching: 4 Acronyms - the missing element in your searching? Online Review, v. 7, n. 6, p. 475-83, Dec. 1983.

SIEVERT, M. E.; BOYCE, B. R. Hedge trimming and the ressurrection of the controlled vocabulary in online searching. Online Review, v. 7, n. 6, p. 489-495, Dec. 1983.

SVENONIOUS, E. Design of controlled vocabularies. In: ENCYCLOPEDIA OF LIBRARY AND INFORMATION SCIENCE. New York : Marcel Dekker, 1988. v. 45, p. .82-108.

_____. The intellectual foundation of information organization. Cambridge : MIT, 2000.

_____. Natural language vs controlled vocabulary. IN: CANADIAN CONFERENCE ON INFORMATION SCIENCE, 1976, Ontario. Proceedings... [S. l. : s. n.], 1976. p. 141-150.

_____. Unanswered questions in the design of controlled vocabularies. JASIS v. 37, n. 5, p. 331-340, 1986.

_____. Use of classification in online retrieval. Library Resources & Technical Services, v. 27, n. 1, p. 76-80, Jan./Mar. 1983.

TENOPIR, C. Searching by controlled vocabulary or free text? Library Journal v. 119, n. 15, p. 58-59, 1987.

_____. Searching Harvard Business Review. Online. v. 9, n. 2, p. 71-78, Mar. 1985.

WAGERS, R. Effective searching in database abstracts. Online, v. 7, n. 5, p. 60-77, Sept. 1983.