Avaliação comparada do uso de linguagens de indexação em catálogos de bibliotecas universitárias para recuperação por assunto

Comparative evaluation of the use of indexing language in catalogs of university libraries for subject retrieval

Vera Regina Casari BOCCATO (1), Mariângela Spotti Lopes FUJITA (2), Isidoro GIL LEIVA (3)

(1) Universidade Federal de São Carlos, Departamento de Ciência de Informação, Brasil, vboccato@ufscar.br.
(2) Universidade Estadual Paulista, Departamento de Ciência de Informação, Brasil, fujita@marilia.unesp.br.
(3) Universidad de Murcia, Facultad de Comunicación y Documentación, España, isgil@um.es

Resumen

Con el objetivo de comprobar la influencia de la disponibilidad de lenguajes de indización en el funcionamiento de los sistemas de recuperación, se realizó un estudio de evaluación comparativa del uso del lenguaje natural y dos lenguajes de indización especializados. El estudio se desarrolló en el ámbito del planteamiento de estrategias de búsqueda por materias en catálogos en línea de bibliotecas universitarias. Para determinar la potencialidad de cada lenguaje de indización en la recuperación por materia se utilizó la fórmula de precisión. Se concluye que en la evaluación comparativa del uso de lenguajes de indización la especificidad de términos requerida por los usuarios en la recuperación fue más satisfactoria realizando la consulta a través de lenguajes controlados, cuya disponibilidad y facilidad también es un requisito imprescindible.

Palabras clave: Lenguajes de indización. Catálogos en línea. Recuperación por materias. Bibliotecas universitarias.

Abstract

A comparative evaluation was made of the use of natural language versus two specialized indexing languages, aiming to demonstrate the influence of the availability of indexing languages on the functioning of information retrieval systems. The study was conducted within the ambit of the construction of search strategies by subject in online university library catalogs. The precision ratio was calculated to determine the accuracy of each indexing language in subject-based information retrieval. From the comparative evaluation of the use of indexing languages, it was concluded that the term specificity required by the user during retrieval was more satisfactory when the query was made through controlled languages, whose availability and simplicity is also an indispensable requisite.

Keywords: Indexing languages. Online catalogs. Subject retrieval. University libraries.

1. Introdução

Com o acesso às informações por meio dos catálogos on-line, mais conhecidos pela sigla OPAC (*Online Public Acess Catalog*) os usuários podem recuperar as informações necessárias por meio de buscas cruzadas em diversos índices, como autor, título, assunto e data.

A recuperação em catálogos on-line traz para os usuários uma lista com registros bibliográficos referentes à pesquisa realizada. Assim, o usuário pode selecionar o registro desejado que vem acompanhado da informação sobre a localização física em formato impresso.

Em catálogos de bibliotecas, os registros bibliográficos são, em sua maioria, de documentos que denominamos de obras avulsas para nos referirmos ao leque de documentos que se apresentam como livros, dissertações, teses, obras de referência e outros, diferente das bibliografias que contêm registros bibliográficos de artigos de periódicos.

A investigação do contexto de tratamento de conteúdo em bibliotecas universitárias é oportuna, considerando-se ser outro ambiente de sistemas de recuperação da informação, cujas funções e usos são diferenciados em razão de objetivos institucionais distintos entre bibliotecas e serviços de informação especializados em áreas de conhecimento. Além disso, a biblioteca é um ambiente institucional muito conhecido por muitos usuários leitores e pesquisadores, cujo interesse é cotidiano.

Catálogos on-line de bibliotecas universitárias são sistemas de recuperação da informação, inseridos em um contexto de áreas científicas especializadas, que necessitam de instrumentos de organização e recuperação da informação compatíveis com a política de indexação da

biblioteca universitária e, também, com a estratégia de busca de sua comunidade usuária.

As linguagens de indexação são instrumentos de organização e recuperação da informação com dupla função, por serem utilizados tanto no processo de indexação quanto na busca da informação em sistemas de recuperação da informação. Na indexação, possibilitam o controle de vocabulário e a representação dos conteúdos dos documentos e, na busca, a compatibilidade da linguagem do usuário para a recuperação das solicitações por assunto dos usuários.

A primeira função da linguagem de indexação é mais evidente para catalogadores, no contexto de atuação profissional de bibliotecas, mas não para os usuários que utilizam o catálogo on-line. Geralmente, são utilizadas listas de cabeçalhos de assuntos que realizam o controle de vocabulário e a representação dos conteúdos dos documentos. A linguagem de indexação é utilizada por catalogadores em consonância com uma política de indexação que estabelece diretrizes para a operação de catalogação de assuntos em bibliotecas e está de acordo com a gestão documental.

Porém, a segunda função da linguagem de indexação, a de atuar na busca por assuntos pelo usuário, depende de sua disponibilização junto à interface de busca do catálogo on-line. O problema na segunda função ocorre pela indisponibilidade da linguagem de indexação para orientar a compatibilização com a linguagem do usuário no momento em que é realizada a estratégia de busca por assuntos no sistema de recuperação da informação.

Poucos catálogos on-line disponibilizam a linguagem de indexação junto à interface de busca por assuntos em catálogos on-line e o usuário usa termos de sua linguagem natural sem nenhum controle de vocabulário gerando buscas mal sucedidas, sem especificidade, precisão, exaustividade e revocação. Isso acontece porque o usuário do catálogo, ao fazer sua busca por assuntos, ignora que existe uma linguagem de indexação que poderá ajudá-lo na escolha de um termo que foi, ao mesmo tempo, escolhido anteriormente pelo catalogador para a representação do conteúdo de determinados documentos.

A indisponibilidade de linguagens de indexação junto às interfaces de busca por assunto poderá provocar sucessivas buscas mal sucedidas pela falta de compatibilidade da linguagem livre usada por usuários com a linguagem de indexação utilizada pelo catalogador. O usuário continuará sua busca com uso de termos mais genéricos para compensar a falta de termos específicos compatíveis até que obtenha uma alta revocação de registros bibliográficos dos documentos da biblioteca que o levará a frustrar-se pela falta de especificidade e precisão. Por este motivo, o uso da busca por assuntos seja, talvez, pouco acionado por usuários porque preferem obter mais precisão na busca por palavras do título e autor.

Essa situação torna-se um verdadeiro círculo vicioso, no qual o catalogador não se preocupa muito com a catalogação de assuntos porque o usuário não realiza a busca por assuntos e o usuário não faz busca por assuntos porque não recupera com a especificidade desejada e também não sabe como obter especificidade e precisão. O rompimento dessa dinâmica pode ser efetuado com a simples disponibilização da linguagem de indexação ao lado da interface de busca.

Resultados de investigações anteriores (Zumer e Zeng, 1994, Anderson, 1998, Miller, 2004, Gross e Taylor, 2005, Fujita, Rubi e Boccato, 2009) indicam mudanças significativas no comportamento informacional de acesso e uso do catálogo por usuários que solicitam recuperação por assuntos com mais especificidade, compatibilidade com sua linguagem de busca e disponibilidade de mecanismos de interação.

Os catalogadores e bibliotecários dirigentes de bibliotecas universitárias precisam se conscientizar da mudança de comportamento informacional de seus usuários e de que existe um problema não resolvido com o não uso das linguagens de indexação na busca por assuntos em catálogos on-line pela falta de disponibilidade e acessibilidade.

Todavia, a conscientização parte de iniciativas de estudos de avaliação de uso de linguagens de indexação e do desempenho do sistema de recuperação por assuntos em catálogos on-line que demonstrem as diferenças, tanto de comportamento informacional quanto de resultados de recuperação com especificidade e precisão. As diferenças servirão de motivação para promover mudanças na política de indexação em bibliotecas, capazes de alterar rotinas e condutas já estabelecidas no contexto de tratamento temático. É preciso convencer, antes de tudo, que a mudança de rotinas trará benefícios ao sistema de recuperação da informação e, por conseguinte, ao uso mais ampliado do conhecimento contido nas coleções da biblioteca.

Com esse propósito, o andamento de duas pesquisas, *Bases científicas e metodologias inovadoras para a interoperabilidade entre linguagens de indexação: uma proposta de investigação para aplicação* e *Política de Indexação para*

Bibliotecas (1), proporcionaram a formação de um grupo com pesquisadores e bibliotecários catalogadores do sistema de bibliotecas da UNESP (2). Em sucessivas reuniões, o grupo trocou e compartilhou experiências práticas e conhecimentos teóricos sobre a compatibilidade entre as linguagens de indexação e a indexação e recuperação de assuntos em seus catálogos, sobre a necessidade do estudo de avaliação comparada do uso de linguagens de indexação e sobre o método de coleta de dados a ser aplicado para a avaliação comparada.

Com o objetivo de verificar a influência da disponibilidade de linguagens de indexação no desempenho do sistema de recuperação realizou-se estudo de avaliação comparada do uso da linguagem natural com duas linguagens de indexação especializadas em estratégias de busca por assuntos em catálogos on-line de bibliotecas universitárias. Para determinar a exatidão de cada linguagem de indexação na recuperação por assuntos foi utilizado o cálculo do índice de precisão.

2. O uso de linguagens de indexação na representação e recuperação da informação científica especializada

Quando nos reportamos ao tema "o uso de linguagens de indexação na representação e recuperação da informação" faz-se necessário consideramos os diversos fatores que norteiam a escolha do instrumento de representação temático mais adequado em sistemas de recuperação da informação de bibliotecas.

No contexto das áreas científicas especializadas, os catálogos on-line de bibliotecas universitárias possibilitam a recuperação da informação por pontos de acessos de autor, título, assunto, entre outros, propiciando ao usuário local e remoto a busca e o uso da informação científica, visando à construção de um novo conhecimento.

Sobre a busca por assunto, esta nos permite conhecer o estado da arte de vários segmentos do conhecimento científico produzido a partir da recuperação de distintos trabalhos elaborados por diferentes autores e épocas, disponibilizados em suportes diversos, possibilitando o delineamento histórico e cultural da temática em estudo, além da identificação de aspectos inovadores, dentro do vanguardismo que a ciência possui.

Nessa perspectiva, os processos e os instrumentos de representação e recuperação da informação tornam-se importantes para o tratamento temático e a disseminação da informação produzida e re-produzida, evidenciando o ciclo da informação e do conhecimento, isto é, conhecimento–informação–conhecimento, numa sucessiva cadeia produtiva do saber.

A catalogação de assunto é o processo de descrever os assuntos tratados em um documento por meio do uso de uma linguagem de indexação. Para Fiúza (1985, p. 257) ela é "[...] a disciplina ou conjunto de disciplinas que tratam da representação, nos catálogos de bibliotecas, dos assuntos contidos no acervo."

Por sua vez, a indexação é "[...] a ação de descrever e identificar um documento de acordo com seu assunto", determinando dois princípios básicos norteadores para o desenvolvimento desse processo: 1) estabelecimento dos conceitos tratados num documento, ou seja, o assunto; e 2) tradução dos conceitos nos termos da linguagem de indexação (UNISIST, 1981, p. 84). A indexação propicia a construção de índices de assunto de bases de dados.

Independentemente das concepções teóricas abordadas, isto é, a catalogação de assunto pela dimensão norte-america e a indexação pela inglesa, tem-se que ambos os processos representam a informação mediante o uso de uma linguagem de indexação.

As diferenças na realização da catalogação de assunto e na indexação são basicamente pautadas "[...] em função do contexto de atuação profissional, dos tipos de suportes em análise e da profundidade com que os processos são efetuados" (Boccato, Fujita, Rubi, 2010, p. 105).

Esses elementos também são subsídios fundamentais para a escolha da linguagem de indexação adotada pelo sistema de recuperação da informação. Tal escolha ocorre entre a linguagem natural e a controlada, e esta última entre um sistema pré-coordenado e um pós-coordenado. As diretrizes estabelecidas pela política de indexação conduzirão a biblioteca para tal decisão que orientará, também, o tratamento temático e o desempenho do sistema em uso.

Recorrendo-se ao estudo de Lopes (2002) vimos que a linguagem natural (linguagem do cotidiano, do discurso comum), nos apresenta vantagens e desvantagens em relação à linguagem controlada. Esta, por sua vez, também nos mostra aspectos favoráveis e não favoráveis em sua aplicação. Como vantagem no uso da linguagem natural tem-se a imediata representação do assunto sem a necessidade de consulta a uma linguagem controlada.

Todavia, ao mesmo tempo em que é efetiva a interação entre o usuário e o sistema na recupe-

ração da informação, ele fará um esforço intelectual maior na identificação de termos ambíguos causados pela polissemia, sinonímia e homonímia próprias da linguagem natural. Isso ocasionará a alta revocação e a baixa precisão no sistema.

Na atualidade temos presenciado o uso da linguagem natural em ambientes informacionais, na *web*. A *web* 2.0 possibilita a interação do usuário na prática da indexação social em redes de informação colaborativas (*blogs, twitters*, entre outros), na formação de um discurso integrado e na modelagem de sistemas sociais de organização do conhecimento, exemplificado pelas *folksonomias*.

A indexação social refere-se "[...] a ação de etiquetagem desempenhada por usuários de ferramentas sociais em ambiente web. (Guedes; Dias, 2010, p. 42).

Sobre as *foksonomias* e na concepção de Wal (2011) elas são o resultado da marcação [etiquetagem/*tagging]* livre e pessoal de informações e objetos (qualquer coisa com uma *URL - Uniform Resource Locator*) para a recuperação dessas mesmas informações e objetos. As *folksonomias* são criadas a partir do ato da etiquetagem realizada pelos usuários que consomem as informações e são feitas em um ambiente social (geralmente compartilhados e abertos aos demais usuários), configurando-se num instrumento social de representação temática.

No que se refere à linguagem controlada, esta permitirá um controle do vocabulário utilizado na representação para a recuperação da informação. O usuário escolherá, dentre o seu repertório terminológico disponibilizado, o termo que melhor representa as suas necessidades informacionais na elaboração de estratégias de busca, tendo em vista a precisão na recuperação.

Todavia, faz-se necessária a disponibilidade da linguagem controlada na interface de busca do sistema, bem como um treinamento sobre o uso que se faz dela e a sua atualização constante de acordo com o desenvolvimento científico e tecnológico.

As listas de cabeçalhos de assunto são linguagens de estruturas pré-coordenadas, constituídas de palavras ou expressões advindas da linguagem natural. Possuem cabeçalhos e subcabeçalhos representativos de diversas áreas do conhecimento, dispostos em uma estrutura hierárquica, relevando o controle de sinônimos e associação entre os cabeçalhos próximos.

Os tesauros são linguagens de indexação controladas, pós-coordenadas, formadas por termos descritores e não descritores advindos das linguagens de especialidade (3) e da linguagem natural, geralmente representativos de uma única área do conhecimento, a partir de relações lógico-semânticas hierárquicas, de equivalência e não-hierárquicas (associativas).

Os tesauros possibilitam a flexibilização na recuperação da informação com o uso de operadores *booleanos* (AND, OR, NOT, entre outros) na combinação de termos, no momento da construção de estratégias, na busca por assunto.

No contexto acadêmico que circunda a produção científica especializada, a universidade tem a tarefa de formar e qualificar pessoas para atuação no meio educacional e técnico-profissional e as bibliotecas universitárias, por sua vez, têm como função contribuir com o desenvolvimento do ensino, da pesquisa e da extensão —os três pilares da universidade, oferecendo serviços e produtos de qualidade, a partir do acesso e busca da informação com rapidez, precisão e pertinência em relação aos desejos de investigação dos usuários locais e remotos.

Os resultados alcançados pela pesquisa de Boccato (2009) revelam que os usuários inseridos num contexto de alta especialização de assunto necessitam de informações relevantes e de acordo com suas necessidades informacionais. Tais informações devem estar disponibilizadas em um catálogo on-line acessível por meio do uso de linguagens controladas, vistas como instrumentos de representação e recuperação da informação mais adequados para tal situação e contexto.

Nesse sentido, este estudo de avaliação comparada entre o uso de linguagem natural, linguagem controlada pré-coordenada e linguagem controlada pós-coordenada torna-se fundamental para que possamos verificar o desempenho e o índice de precisão na recuperação de cada uma delas em catálogos coletivos on-line de áreas científicas especializadas de bibliotecas universitárias.

3. Metodologia

Os procedimentos metodológicos referem-se ao desenvolvimento da avaliação comparada de estratégias de buscas por assunto em catálogos on-line a partir do uso de Linguagem Natural (LN) e de duas linguagens controladas: a Lista de Cabeçalhos de Assunto da Rede BIBLIODATA (LCARB), utilizada no catálogo coletivo ATHENA da Universidade Estadual Paulista (Unesp) e o Vocabulário Controlado do SIBi/USP (VocaUSP) utilizada no catálogo coletivo

DEDALUS da Universidade de São Paulo (USP).

O uso das linguagens natural e controlada na estratégia de busca para recuperação por assuntos no catálogo on-line ATHENA da UNESP foi realizado pelos usuários em três etapas na coleta de dados: 1) uso da linguagem natural; 2) uso da linguagem de indexação da UNESP, o BIBLIODATA (LCARB); 3) uso da linguagem de indexação da USP, o Vocabulário Controlado do SIBI/USP (VocaUSP).

A análise dos dados coletados ocorreu mediante a aplicação do índice de precisão da recuperação da informação.

3.1. O universo e os sujeitos do estudo

O universo deste estudo foram sete bibliotecas do sistema de Bibliotecas da Unesp, Brasil:

Área	*Campus*	*Curso*	*Sigla*
Ciências Agrárias	Dracena	Zootecnia	BDR
Ciências Exatas	Bauru	Ciência da Computação	BBA
Ciências Biológicas	Botucatu Araçatuba Araraquara	Med.Veterin. Odontologia Nutrição	BBO BFO BFA
Ciências Humanas	Marília Assis	C. Sociais Letras, Hist.	BMA BAS

Quadro I. Bibliotecas do universo de estudo por áreas do conhecimento

Os objetos de estudo empíricos foram a linguagem natural, a Lista de Cabeçalhos de Assunto da Rede BIBLIODATA e o Vocabulário Controlado do SIBi/USP, Brasil.

A linguagem natural foi caracterizada pela linguagem que o usuário utilizou no momento da realização da busca por assunto. A Lista de Cabeçalhos de Assunto da Rede BIBLIODATA, elaborada pela Fundação Getúlio Vargas (s. d.), é a linguagem de indexação utilizada para a indexação e recuperação da informação no catálogo coletivo da Unesp–ATHENA que integra os acervos das trinta e quatro bibliotecas da Rede Unesp. A LCARB é uma linguagem pré-coordenada, de cabeçalhos de assunto autorizados e não autorizados. Os cabeçalhos de assuntos são compostos por cabeçalhos e subcabeçalhos de diversas áreas do conhecimento, previamente estabelecidos no momento da indexação dos documentos. A Lista de Cabeçalhos de Assunto da Rede BIBLIODATA não é disponibilizada pelo sistema de recuperação da informação na busca por assunto.

O Vocabulário Controlado do SIBi/USP (2006-8) foi desenvolvido pelo Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBi/USP). O VocaUSP é uma linguagem pós-coordenada, constituída de um repertório terminológico formado a partir da linguagem de especialidade e da linguagem natural de termos descritores e não descritores de várias áreas do conhecimento. Os relacionamentos hierárquicos e de equivalência existentes entre os termos constituem a macroestrutura do VocaUSP, possibilitando a representação e a recuperação por assunto dos documentos no catálogo coletivo da USP–DEDALUS que agrega o acervo das quarenta e quatro bibliotecas do SIBI/USP. O usuário do catálogo DEDALUS tem acesso ao Vocabulário Controlado do SIBi/USP no momento da elaboração da estratégia de assunto para a recuperação da informação.

Os sujeitos deste estudo foram os usuários discentes de graduação que desenvolvem pesquisas de Iniciação Científica (IC) com apoio de agências de fomento brasileiras (4), discentes de pós-graduação em nível de mestrado e doutorado, totalizando vinte e um participantes. Todos os usuários realizaram as buscas por assuntos com uso da linguagem natural, LCARB e VocaUSP, no catálogo ATHENA, conforme descrito, a seguir.

3.2. Procedimentos metodológicos

1) Fase inicial:

- apresentação a todos os usuários do objetivo da pesquisa;
- realização da familiarização sobre o uso do catálogo ATHENA e das linguagens de indexação LCARB e VocaUSP;
    - realização de familiarização sobre o roteiro da pesquisa por assunto: uso da linguagem natural a partir de palavras-chave previamente escolhidas e de acordo com o assunto da pesquisa em desenvolvimento, para a elaboração da estratégia de busca a ser realizada no catálogo ATHENA;
    - consulta à LCARB disponibilizada em CD-ROM a partir das palavras-chave previamente escolhidas e de acordo com o assunto da pesquisa em desenvolvimento, identificando e anotando o termo correspondente (sem o hífen quando houver) (5) para a elaboração

da estratégia de busca a ser realizada no catálogo ATHENA;

- consulta ao VocaUSP a partir das palavras-chave previamente escolhidas e de acordo com o assunto da pesquisa em desenvolvimento, identificando e anotando o termo correspondente para a elaboração da estratégia de busca a ser realizada no catálogo ATHENA;

- solicitação aos usuários para selecionarem os documentos relevantes recuperados de acordo com suas necessidades de informação;
- instalação do *software Free Screen to Video 1.2* (6) para a captura das telas de busca//gravação realizadas pelos discentes.

2) Fase do desenvolvimento:

- solicitação aos usuários que iniciem as buscas e que não realizem nenhuma intervenção dialógica ou de comportamento facial ou corporal com as pesquisadoras;
- iniciar a gravação das telas de busca realizadas pelos usuários.

3) Fase final:

- exportação aos computadores das respectivas bibliotecas participantes do arquivo gerado mediante a gravação das telas de busca realizadas pelos usuários;
- importação para o computador das pesquisadoras do arquivo gerado mediante a gravação das telas de busca realizadas pelos usuários;
- leitura das telas gravadas mediante o *software Free Screen to Video 1.2,* na localização e anotação do número total de documentos recuperados e dos termos utilizados advindos da linguagem natural, da LCARB e do VocaUSP, no catálogo ATHENA, por pesquisa e por usuário, de cada biblioteca;
- verificação das anotações realizadas sobre o número de documentos relevantes recuperados, a partir dos julgamentos realizados pelos usuários, por pesquisa, de cada biblioteca a partir do uso da linguagem natural, da LCARB e do VocaUSP, no catálogo ATHENA;
- elaboração de quadro de registros, em *excel*, intitulado "Coleta de dados geral quantitativa final" (vide Apêndice A) de todas as buscas realizadas com o uso da linguagem natural, da LCARB e do VocaUSP no catálogo ATHENA, contendo, respectivamente as palavras-chave escolhidas, os termos encontrados nas respectivas linguagens de indexação e utilizados nas buscas, a quantidade dos documentos recuperados e a quantidade dos documentos relevantes recuperados, em cada catálogo;
- análise dos dados coletados, a partir do quadro registro e das telas gravadas, visando à comparação entre linguagens por meio da determinação do índice de precisão, conforme exemplos abaixo demonstrados extraídos do Apêndice A:

Buscas por	*Termos*	*Recup.*	*Relev.*
Ling.Natural	mecanização	124	6
LCARB	Máquinas agrícola— manutenção e reparo	5	1
VocaUSP	Equipamentos agrícolas	2	0

Quadro II. Biblioteca BDR: Zootecnia

Buscas por	*Termos*	*Recup.*	*Relev.*
Ling.Natural	Soberania alimentar	2	0
LCARB	Teoria crítica AND Relações Internacionais	7	2
VocaUSP	Teoria crítica AND Relações Internacionais	7	2

Quadro III. Biblioteca BMA: Ciências Sociais

Buscas por	*Termos*	*Recup.*	*Relev.*
Ling.Natural	Metrologia	26	8
LCARB	Metrologia	26	8
VocaUSP	Metrologia	26	8

Quadro IV. Biblioteca BBA: Ciência da Computação Biblioteca

Buscas por	*Termos*	*Recup.*	*Relev.*
Ling.Natural	Úlcera duodenal	1	1
LCARB	Ulcera péptica	7	4
VocaUSP	Absorção (fisiologia)	17	0

Quadro V. BBO: Medicina Veterinária

Foram realizadas 248 buscas com 21 usuários em 7 bibliotecas (média de 3 usuários em cada biblioteca). O total de documentos recuperados e documentos relevantes em todas as buscas realizadas por Linguagem Natural e Linguagens de indexação LCARB e VocaUSP foi de:

Linguagem	*Total*	*Recuperados*	*Relevantes*
Ling.Natural	139	5.039	366
LCARB	69	4.552	298
VocaUSP	40	2.086	163
Total	248	11.677	827

Quadro VI. Buscas com uso de linguagens: documentos recuperados e documentos relevantes

3.3. Análise dos dados

A análise dos dados foi realizada mediante a adoção do índice de precisão da recuperação da informação dos termos utilizados na estratégia de busca.

De acordo com os pressupostos metodológicos de Lancaster (2004, p. 4) o índice de precisão (*precision*) é determinado a partir da relação existente entre número de documentos relevantes (7) recuperados e o número de total de documentos recuperados pelo sistema.

A precisão na recuperação da informação ocorre a partir da submissão do resultado de uma busca ao usuário na qual ele seleciona um conjunto de documentos que considera relevante em relação a sua necessidade de pesquisa inicial. Para Lancaster (2004), a precisão é a capacidade de evitar documentos inúteis na recuperação pelo sistema.

O índice de precisão (*Pre*.) pode ser determinado a partir da seguinte fórmula:

$$Pre. = \frac{nº\ de\ itens\ relevantes\ recuperados\ (a)}{nº\ total\ de\ itens\ recuperados\ (b)}$$

Onde, (a): itens (documentos) relevantes recuperados a partir do julgamento do usuário; (b): itens (documentos) recuperados pelo sistema de informação.

Nesta pesquisa, utilizamos os seguintes procedimentos para a determinação do índice de precisão:

- consulta ao quadro de registros "Coleta de dados geral quantitativa final" para a realização da somatória do número de documentos relevantes recuperados, por pesquisa, por usuário, de cada biblioteca, com a utilização da linguagem natural, da LCARB e do VocaUSP, no catálogo ATHENA;
- consulta ao quadro de registros "Coleta de dados geral quantitativa final" para a realização da somatória do número total de documentos recuperados, por pesquisa, por usuário, de cada biblioteca, mediante o uso da linguagem natural, da LCARB e do VocaUSP, no catálogo ATHENA;
- determinação do índice de precisão do uso da linguagem natural, da LCARB e do VocaUSP, no catálogo ATHENA, por buscas realizadas pelos usuários.

4. Resultados

Os resultados da avaliação comparada do uso de linguagem de indexação demonstraram que a especificidade de termos exigidas pelo usuário na recuperação não apresentou muita diferença entre a Linguagem Natural e as Linguagens LCARB e VocaUSP. Na comparação dos índices de precisão evidenciou-se mais especificidade no VocaUSP (7,8%) em relação à Linguagem Natural (7,2%) e LCARB (6,5%).

$$Pre. = \frac{366}{5039} = 0{,}072 = 7{,}2\%$$

Figura 1. Índice de precisão da recuperação por assunto do catálogo ATHENA com o uso da linguagem natural.

$$Pre. = \frac{298}{4552} = 0{,}065 = 6{,}5\%$$

Figura 2. Índice de precisão da recuperação por assunto do catálogo ATHENA com o uso da LCARB.

$$Pre. = \frac{163}{2086} = 0{,}078 = 7{,}8\%$$

Figura 3. Índice de precisão da recuperação por assunto do catálogo ATHENA com o uso do Vocabulário USP.

O VocaUSP foi construído a partir da linguagem de especialidade e da linguagem do usuário abrangendo várias áreas científicas especializadas, provido de termos genéricos e específicos, organizados hierarquicamente e contemplando o controle da ambiguidade e de sinônimos, características essas provenientes da linguagem natural. A macroestrutura permite ao usuário navegar nas categorias e subcategorias ampliando ou restringindo a busca de acordo com sua necessidade informacional. A pós-coordenação entre os termos possibilita tal situação por meio da combinação entre eles no momento da elaboração da estratégia para a busca por assunto e recuperação precisa da informação.

Por outro lado, a LCARB alcançou um índice de precisão no catálogo ATHENA diferente do VocaUSP, isto é, de 6,5%, em relação a 7,8%. Atribuímos esse declínio de quase 2% ao fato da linguagem possuir uma estrutura pré-coordenada que não propicia uma flexibilização

na construção de estratégias de busca de acordo com a intenção de pesquisa do usuário.

Recorrendo ao estudo de Boccato (2009) em que realizou a avaliação do uso da LCARB no catálogo ATHENA pela perspectiva sociocognitiva do indexador e do usuário, destacamos dentre os resultados obtidos, a falta de vocabulário de especialidade e da indisponibilidade e inacessibilidade da linguagem no sistema. Primeiramente, podemos considerar que a falta de repertório de especialidade é pelo fato de uma lista de cabeçalhos de assunto ser modelada a partir da linguagem natural, não contemplando, em sua maioria, vocabulários de especialidade representativos de áreas científicas. Além disso, as listas de cabeçalhos de assunto privilegiam mais a generalidade do que a especificidade de cabeçalhos e subcabeçalhos, não promovendo, plenamente, a exatidão na recuperação da informação.

Um aspecto também observado neste estudo foi a satisfação do usuário no acesso à linguagem a partir da consulta e uso, em CD-ROM, na construção da estratégia de busca. A LCARB não é disponibilizada pelo catálogo ATHENA e, consequentemente, não é acessível ao usuário na representação, por cabeçalhos, de suas questões de pesquisas. Isso é um aspecto de suma importância, pois a linguagem de indexação deve estar ao alcance do usuário para que a busca por assunto possa ser realizada a partir da linguagem e com a linguagem. Essa disponibilização encontramos no VocaUSP, bem como em outros sistemas de organização do conhecimento brasileiros de áreas científicas especializadas como o DeCS —Descritores em Ciências da Saúde*,* elaborada pela BIREME— Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (2011).

Na mesma configuração da LCARB, faz-se presente a *Terminologia de Assunto*, lista de cabeçalhos de assunto utilizada na indexação e recuperação da informação no catálogo da Fundação Biblioteca Nacional (Brasil) (s. d.), propiciando ao usuário o acesso aos cabeçalhos e subcabeçalhos que formam o seu repertório terminológico de diferentes áreas do conhecimento.

Numa análise sistematizada por áreas do conhecimento, as Ciências Biológicas (Medicina Veterinária, Odontologia, Nutrição) e as Ciências Humanas e Sociais (Letras, História e Ciências Sociais) mostraram um desempenho melhor no uso do VocaUSP, respectivamente, 13% e 12%. Nas Ciências Exatas (Ciência da Computação) houve uma preferência para a LCARB com um índice de precisão de 7%. As Ciências Agrárias (Zootecnia) apontaram um desempenho similar ocorrido entre o VocaUSP e a LCARB, atingindo um índice de 5% e 6%, respectivamente. A linguagem natural obteve bom desempenho em todas as áreas (CA – 7%; CE-9%; CHS-13%) com exceção de Ciências Biológicas (5%).

Sobre isso, verificamos que a LCARB utilizada no catálogo ATHENA teve um índice de precisão na recuperação da informação mais baixo em relação às outras duas linguagens, perfazendo 6,5% em relação a 7,8% alcançado pelo VocaUSP e 7,2% pela Linguagem Natural (Gráfico I).

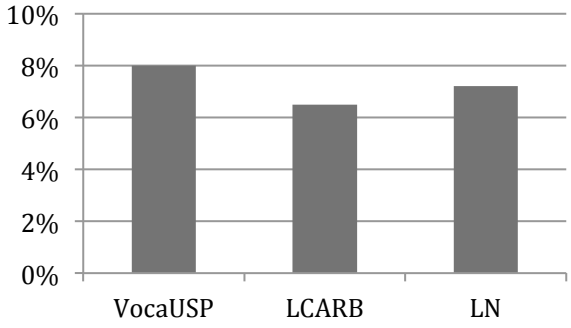

Gráfico I. Índice total de precisão das linguagens de indexação na recuperação por assunto no catálogo ATHENA.

Os resultados entre as três linguagens apresentaram poucas diferenças, possivelmente devido à realização de estratégias de buscas em modo simples e não avançado no catálogo online, o que demonstra a necessidade de continuidade dessa investigação.

5. Considerações finais

Na contemporaneidade, assistimos ao uso da linguagem natural em espaços colaborativos de informação como as redes sociais, *blogs*, entre outros que permitem a prática da indexação social para a recuperação da informação. Tal aplicabilidade é pertinente, pois o universo informacional em que permeia o uso da linguagem natural é proveniente de assuntos diversificados, muitas vezes, gerais e utilizado por diferentes categorias de usuários que requerem um processo de comunicação dinâmico, interativo, entre o usuário e o ambiente de informação, sem a preocupação causada pela baixa precisão na recuperação que a linguagem natural promove.

Por outro lado, numa unidade de informação de cunho científico, caracterizada pelas bibliotecas universitárias, os usuários —docentes e/ou pesquisador, discente de graduação e de pós-graduação— exigem uma especificidade no tratamento do assunto, com vistas à precisão na recuperação da informação especializada. Tais usuários, por sua vez, solicitam realizar a busca por assunto facilitada por uma linguagem controlada que permite o controle do vocabulário, inibindo a presença da polissemia, sinonímia e homonímia causadas pela ambigüidade da linguagem natural.

Nessa perspectiva, vimos que espaços informacionais distintos necessitam de tratamentos e uso de linguagens diferenciadas. Subsidiado por Boccato, Fujita e Rubi (2010) a escolha dessa linguagem deve levar em conta o ambiente e o sistema de informação, a categoria de usuário, o tipo de suporte, bem como o grau de profundidade em que o documento é tratado. No contexto das bibliotecas universitárias, acreditamos ser a linguagem controlada a mais indicada, considerando-se o catálogo que disponibiliza documentos impressos e eletrônicos, o usuário especialista e o documento, independentemente do suporte, que deve ser representado o mais específico possível. Tal linguagem deve estar disponibilizada no catálogo on-line para o acesso e uso em estratégias de busca para a recuperação precisa da informação.

Agradecimentos

A coleta dos dados da pesquisa foi orientada pelos autores/pesquisadores, porém executada pelos bibliotecários catalogadores do sistema de bibliotecas da UNESP que compõem o grupo de pesquisa para o desenvolvimento da política de indexação, motivo pelo qual agradecemos a colaboração inestimável desse envolvimento autorizado pela Coordenadoria Geral de Bibliotecas da UNESP.

Os cálculos realizados para obtenção dos índices de precisão tiveram a colaboração da mestranda Mariana de Oliveira Inácio, do Programa de Pós-Graduação em Ciência da Informação, a quem agradecemos a dedicação.

Notas

(1) Pesquisas financiadas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil com pareceres favoráveis dos Comitês de Ética da FFC/UNESP e UFSCar para a coleta de dados.

(2) Bibliotecários catalogadores: Cláudio Hideo Matsumoto (Araçatuba), Fábio Sampaio Rosas (Dracena), Maria Marlene Zaniboni (Bauru), Sô-nia Scutari (Araraquara/Farmácia), Sulamita Selma Clemente Colnago (Botucatu), Telma Ja-queline Dias Silveira (Marília), Vânia Ap. Mar-ques Favato (Assis) e Cássia Adriana de Sant'Ana Gatti (CGB/Marília)

(3) A linguagem de especialidade: "[...] utilizada pelo pesquisador na geração do conhecimento, proveniente das atividades desenvolvidas em grupos de pesquisa e/ou no momento da reali-zação do seu discurso científico (termos especi-alizados constantes nos trabalhos científicos como os artigos de periódicos etc.)." (Boccato, 2009, p. 119).

(4) PIBIC/CNPq: Programa Institucional de Bolsas de Iniciação Científica do Conselho Nacional de Desenvolvimento Científico e Tecnológico e FAPESP: Fundação de Amparo à Pesquisa do estado de São Paulo, Brasil;

(5) Máquinas agrícolas—Manutenção e reparo. Utilizar sem o hífen: Máquinas agrícolas Manu-tenção e reparo.

(6) Free Screen to Video 1.2: software de captura de tela disponível gratuitamente na internet.

(7) Lancaster (2004) considera como equivalentes as expressões úteis, relevantes e pertinentes.

Referencias

Anderson, S. (1998). A new horizon: an evaluation of a library online public access catalogue. // Library & Information Research News. ISSN 0141-6561. 22:72 (1998) 15-24.

BIREME (2011). DeCS – Descritores em Ciências da Saúde. http://decs.bvs.br/

Boccato, V. R. C. (2009). A linguagem de indexação vista pelo conteúdo, forma e uso na perspectiva de catalogadores e usuários. // Fujita, M. S. L. (org.). A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias: Um estudo de observação do contexto sociocognitivo com protocolos ver-bais. São Paulo: Cultura Acadêmica. ISBN 978-85-7983-015-0. Cap. 6, p. 119-35. http://www.culturaacademica.com.br/titulo_view.asp?ID=56(2010-06-12).

Boccato, V. R. C.; Fujita, M. S. L. Rubi, M. P. (2010). Estudio observacional del contexto sociocognitivo de la catalogación de matérias em bibliotecas universitarias. // Scire: representación y organización del conocimiento. ISSN 1135-3616. 16:2 (jul./dic 2010) 103-115.

Fiúza, M. M. (1985). O ensino da catalogação de assunto. // Revista da Escola de Biblioteconomia da UFMG. ISSN. 0100-0829. 14:2 (set. 1985) 257-269.

Fujita, M. S. L.; Rubi, M. P.; Boccato, V. R. C. (2009) O contexto sociocognitivo docatalogador em bibliotecas universitárias: perspectivas para uma política de tratamento da informação documentária. // DataGramaZero: Revista de Ciência da Informação. ISSN 1517-380. 10:2 (abr. 2009). http://www.datagramazero.org.br. (2011-04-09).

Fundação Biblioteca Nacional (Brasil) (s. d). Terminologia de Assuntos. http://www.bn.br/site/pages/catalogos/terminologiaAssuntos/content.htm.

Fundação Getúlio Vargas (s. d.). Lista de Cabeça-lhos de Assunto da Rede BIBLIODATA. http://www8.fgv.br/biblio data/site2/pesquisa/frame_pesquisa.asp?msg. (Acesso restrito).

Gross, T., Taylor, A. G. (2005). What have we got to lose? the effect of controlled vocabulary on keyword searching results. // College & Research Libraries. ISSN 0010-0870. 66:3 (May 2005) 212-230.

Guedes, R. de M.; Dias, E. J. W. (2010). Indexação social: abordagem conceitual. // Revista ACB: Biblioteconomia em Santa Catarina. ISSN 1414-0594.15:1 (jan./jun. 2010) 39-53.

Lancaster, F. W. (2004). Indexação e resumos: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. 2. ed. Brasília: Briquet de Lemos, 2004. ISBN 85-8563724=2.

Lopes, I. L. (2002). Uso das linguagens controlada e natural em bases de dados: revisão da literatura. // Ciência da In-formação. ISSN 1518-8353. 31:1 (jan./abr. 2002) 41-52.

Miller, D. H. (2004). User perception and the online catalogue: public library OPAC users "think aloud". // McIlwaine, I. A.C. (ed.). Knowledge organization and the global information society: proceedings of the 8th International ISKO Conference, 13-16 July 2004, London, UK. London: Ergon Verlag, 2004. ISBN 3-89913-357-9. 9, 275-280.

USP, SIB (2006-2008). Vocabulário Controlado do SIBi/USP. São Paulo: Universidade de São Paulo. http://143.107.73.99/Vocab/Sibix652.dll

UNISIST. (1981). Princípios de indexação. Revista da Escola de Biblioteconomia da UFMG. ISSN 0100-0829. 10:1 (mar. 1981) 83-94.

Wal, T, V. (2011). Folksonomy: coinage and definition. http://www.vanderwal.net/folksonomy.html (2011-04-04).

Zumer, M.; Zeng, L. (1994). Comparison and evaluation of OPAC end-user interfaces. Cataloging & Classification Quarterly. // ISSN 0163-9374. 19:2 (1994) 67-98.

Apêndice A: Quadro de registros "Coleta de dados geral quantitativa final"

Áreas de conhecimento	*Termos de busca*	*Recuperados na base de dados Athena*	*Relevantes para as necessidades de informação*
Ciências Agrárias (Agronomia)	Implementos agrícolas	8	1
Ciências Biológicas (Odontologia)	Odontologia—Estudo e ensino	21	7
Ciências Exatas (Ciência da Computação)	Redes Neurais Artificiais	22	2
Ciências Humanas (Filosofia)	Estética Kant	15	9

Quadro VII. Coleta de dados geral quantitativa final