TESAURO DE CIÊNCIA DA INFORMAÇÃO EM LÍNGUA PORTUGUESA: desenvolvimento de metodologia

Else Benetti Marques Válio[1]
Cecília C. Cunha Pontes[2]

INTRODUÇÃO

O desenvolvimento de metodologia com o objetivo de contribuir para a construção de **Tesauro em Ciência da Informação em Língua Portuguesa** desvela os princípios básicos dessa área, que podem ser traduzidos por conteúdos textuais em termos documentários. Esses termos, representativos das informações contidas no texto, são identificados por unidades lexicais da língua em apreço que se transformam em linguagem documentária, circunscrita aos traços sintáticos e semânticos, definidos em descritores (CINTRA, 1983).

Entende-se, nesse enfoque, que a técnica mais adequada para a indexação é aquela que possibilita o armazenamento da informação e o questionamento do usuário diretamente em linguagem natural. Essa possibilidade vem sendo viabilizada à medida em que os progressos das pesquisas na área de inteligência artificial estão sendo aperfeiçoados. Por enquanto, a utilização de técnicas automatizadas de indexação, tendo em vista os métodos linguísticos e estatísticos, têm sido considerados eficientes (ANDREEWSKI & RUAS, 1983), uma vez que procuram aproximar-se o mais possível da linguagem natural.

Isto não significa que a indexação automática seja mais eficiente do que a manual. O que se pretende afirmar é que no caso da primeira já existem técnicas que estão mais próximas da busca e recuperação da informação em linguagem natural. As primeiras tentativas para a indexação e elaboração de resumos baseavam-se em contagem de palavras. Em seguida foram introduzidas quatro medidas de significância, fundamentadas em:

- palavras-chave, que eram determinadas estatisticamente;
- palavras-chave características ("clue words"), indicando a inteção do autor em enfatizar um enunciado;
- palavras de títulos e cabeçalhos, com um suposto peso particular;
- posição da frase na estrutura geral. Técnica "autostrating", que se baseia essencialmente no conteúdo do documento, dele podendo apenas extrair frases (FOSKETT, 1973).

As tendências atuais com respeito à indexação automática tratam de abordagens conceituais que estão vinculadas ao estudo linguístico computacional, baseado na eliminação do texto das palavras vazias de significado. Essa técnica, já testada com sucesso, define um sistema versátil de indexação automática de textos - o sistema AUTOMINDEX, por exemplo - que é um subsistema do sistema BIB/DIALOG (ROBREDO, 1991).

Na busca da técnica mais adequada, o pesquisador vai considerar o modo de comportamento dos usuários, frente a recuperação, com vistas a afinidade semântica entre a linguagem documentária, a competência linguística do usuário e o conteúdo lingüístico do documento indexado (GUTIERREZ, 1989).

Este laço semântico latente na representação formal dos conceitos, necessita coincidir com a possibilidade de comunicação entre o texto original e as informações desejadas pelo usuário no momento da recuperação.

Tratando-se da indexação manual, há interferência de variáveis como:

- o objetivo do sistema de recuperação;
- as limitações da linguagem adotada;
- a política de indexação;
- a natureza da área de conhecimento;
- a natureza e propriedade da informação científica e tecnológica.

A interferência dessas variáveis pode ser entendida como subjetivismo e inconsistência do trabalho do indexador. No entanto, o cuidado na escolha das técnicas poderá amenizar a intervenção tanto dessas variáveis como do subjetivismo do especialista da informação, Essa questão já foi estudada por PINHEIRO (1978), com o objetivo de medir a consistência existente em um grupo de indexadores, através do grau de concordância ou discordância na escolha de um termo ou de um conjunto de termos.

Com o surgimento do KWIC, no qual a indexação tornou-se mais elaborada, ao utilizar técnicas de pesquisas mediante o estudo do funcionamento das estruturas sintática e semântica do texto, e com custo

1. Professora Titular do Departamento de Pós-Graduação em Biblioteconomia da Pontifícia Universidade Católica de Campinas
2. Professora Titular do Departamento de Pós-Graduação em Biblioteconomia da Pontifícia Universidade Católica de Campinas

mais baixo, surgiu um novo tipo de profissional da informação que, segundo GOMES (1989), é ainda um profissional sem denominação.

Desse modo, a pesquisa terminológica exige procedimentos anteriores à elaboração de um tesauro, tais como, levantamento de termos, estruturação de conceitos, análise de aspectos lingüísticos sintático-semântico, assim como um minucioso estudo sobre a operacionalização de outros tipos de tesauros (GOMES, 1989).

ELABORAÇÃO DE TESAUROS: estudo de metodologias

Ao relatar a experiência de construção de um **Tesauro em Ciência da Informação**, ARAÚJO (1986) afirma que para iniciar a elaboração são necessárias duas decisões básicas. Primeiro, se é preciso construir um novo tesauro ou se a adaptação de uma linguagem já existente cumpre o objetivo. Segundo, qual é preferível: uma linguagem controlada ou uma linguagem livre?

Analisando esses procedimentos, a Autora apresenta as vantagens de um tesauro em linguagem livre, argumentando que se o sistema for mecanizado a relação custo-efetividade poderá ser mais interessante para a indexação e a busca. O uso da linguagem livre evita distorções como aquelas que ocorrem durante a tradução da linguagem natural para os padrões de uma linguagem de indexação. Para a extração das palavras poderá ser utilizado o texto no todo, o resumo ou o título do documento, automaticamente ou selecionadas manualmente pelo indexador.

Outras vantagens no uso da linguagem livre, afirma a Autora, estão relacionadas à ausência de erros de indexação e nenhuma possibilidade de perda de especificidade. Acrescenta ainda que não há demora na incorporação de novos **termos** no vocabulário, pois estes serão imediatamente incluídos na linguagem assim que aparecer no texto, resumo ou título e que é similar em estrutura ao tesauro de linguagem controlada.

Considerações também são feitas com relação à estrutura escolhida na elaboração do tesauro, tendo em vista a necessidade de fornecer sinônimos, quase sinônimos e todas as alternativas possíveis na grafia das palavras, assim como lidar com conceitos formados por diferentes palavras, incluindo-os no sistema, quer propriamente como conceitos, quer como termos de entrada. As relações hierárquicas assim como outras entre os termos foram orgaizadas alfabética e sistematicamente, objetivando poupar o esforço intelectual do usuário, quando utiliza o tesauro.

São citados ainda na estrutura desse tesauro os artíficos especiais disponíveis que auxiliam na busca, tais como: a) artifício de renovação - "TRUNCAGEM" - as palavras são truncadas e as raízes são usadas para ampliar a busca; b) artifício de precisão - "PESOS" e "AUXILIARES" de ocorrência - o técnico de busca pode utilizá-los para solicitar que certas palavras apareçam juntas no texto, no mesmo parágrafo ou frase.

Entretanto pode-se dizer que o tesauro em linguagem livre traz a desvantagem do esforço, no momento de busca, em fazer coincidir as palavras do documento com as palavras do usuário.

O **Tesauro en Documentación y Información** (RIOS & HERRAN, 1980) foi criado em atenção aos pedidos formulados em vários congressos pela necessidade de uma linguagem de indexação em língua espanhola.

Composto por 560 termos, esse tesauro está organizado em ordem alfabética por descritores e não descritores, nas línguas espanhol-inglês e inglês-espanhol, seguido de glossário, no qual é apresentada uma listagem de definições de todos os termos incluídos.

Quanto à metodologia, a estruturação do tesauro parte da seleção de palavras em linguagem natural, utilizando-se do método indutivo e do método dedutivo. Para a análise de 200 documentos da literatura especializada disponível, com o objetivo de constituir o vocabulário inicial, foi utilizado o método indutivo. Para a complementação desse vocabulário, foi realizada uma revisão de literatura sobre a área. Em seguida, através do método dedutivo, foram consultados os especialistas das áreas envolvidadas e feita a comparação com a Terminologia de Documentação da UNESCO.

A estruturação temática do **Tesauro de Ciência de Informação: versão preliminar** (IBICT, 1989) foi definida em 7 categorias: Informação, Documento, Unidade de Informação, Planejamento, Organização e Administração de Unidades de Informação, Processos e Serviços de Informação, Transferência e Uso da Informação e Profissão.

Inicialmente foi usada para a coleta de termos uma versão preliminar do Macrotesauro de Ciência de Informação e um vocabulário controlado do Centro de Documentação e Informação do IBICT, como instrumento complementar ao Macrotesauro, pois este não atendia as exigências de especificidade necessárias à indexação dos documentos da área. Além desses dois instrumentos também serviram como fontes de pesquisa para a seleção de 972 termos os documentos: FID (1987); LIBRARY & INFORMATION SCIENCE ABSTRACTS (1985); TERMINOLOGY OF DOCUMENTATION (1976); UNESCO thesaurus (1977); RESUMOS DE INFORMAÇÃO (1986) e CIÊNCIA DA INFORMAÇÃO (1985).

A estruturação geral do tesauro compreendeu uma seleção prévia dos termos que, após análise, foram considerados eleitos como descritores e não

descritores, em uma relação de equivalência e de hierarquia estabelecidas a partir da categorização.

O levantamento e a análise de pesquisas sobre o assunto e o estudo de metodologias utilizadas na elaboração de tesauros contribuiram para a definição do material e do método de pesquisa, que possibilitaram atingir ao objetivo de Contribuir para a construção de um vocabulário em Límgua Portuguesa em Ciência da Informação.

Material e Método

Com caráter documental e experimental, o objeto de análise para o desenvolvimento de uma metodologia, visando a elaboração de tesauro, são as dissertações/teses defendidas nos cursos de Pós-Graduação em Biblioteconomia e Ciência da Informação (de 1970 a 1991), existentes no Brasil, em um total de 420 documentos, por constituir-se no conjunto mais representativo das pesquisas realizadas em Ciência de Informação.

Para o processamento e organização das informações significativas dos documentos, está sendo gerada uma base de dados em MICRO-ISIS, das dissertações/teses defendidas nos seis Programas de PG, contendo os título, os resumos e as referências bibliográficas dos documentos.

O trabalho foi iniciado pela indexação manual, cujos parâmetros servirão de modelo para a indexação automática. Após proceder-se a indexação manual das dissertações/teses, deverá ser solicitado ao autor que indique os termos significativos para a identificação e seleção de seu trabalho, a partir das listas conseguidas dos termos candidatos a descritores.

A indexação automática do documento está sendo realizada através de técnicas voltadas para as análises sintática e semântica dos títulos e referências bibliográficas dos documentos/objeto da pesquisa, utilizando-se de método linguístico e estatístico. Para tanto, estão sendo realizados estudos para identificar um **software** adequado. Caso não exista, será desenvolvido um programa próprio às necessidades.

Na análise de conteúdo, está sendo utilizado o método indutivo, a partir de adaptação da proposta de CINTRA (1983), que tem como estudo os núcleos semânticos. Esses núcleos semânticos são organizados "dentro do vocabulário especializado da área, que de si já constitui um campo semântico" e distribuídos em "classes a partir do seguinte raciocínio: a área lida com um **objeto**, apresentado em determinados **suportes materiais**, submetidos a determinadas **operações**, para certos fins" (CINTRA, 1983, p.15). Enfocando esses quatro pólos os termos estão sendo coletados e distribuídos em sete categorias semânticas: objeto, lugar, agente, modo, instrumento, produto e finalidade.

A estratégia de coleta de dados (indexação manual), tendo em vista o levantamento de termos significativos e não significativos, tem como campo de busca as palavras e/ou unidades lexicais constitutivas dos títulos dos documentos - objetos da pesquisa-, dos resumos e títulos das referências bibliográficas das dissertações/teses.

A busca de termos teve como pré-teste a análise de conteúdo de um documento que não fazia parte do acervo escolhido como objeto de estudo. Após a análise e levantamento dos termos, os mesmos procedimentos foram aplicados a cinco dissertações/tese escolhidas.

Extraídos os termos do título, do resumo e das referências bibliográficas em Português, procedeu-se à junção dos termos, levantados nos seis documentos, por categorias semânticas. Eliminados os termos que não apresentavam consistência conceitual para serem descritores, realizou-se a conferência daqueles eleitos com a terminologia do **TESAURO**

(IBICT, 1989). Após essas análises, foram geradas listagens dos termos candidatos a descritores.

Posteriormente também os termos conseguidos deverão ser comparados com aqueles existentes na obra da UNESCO - **Terminology of Documentation**.

Resultados Preliminares

O banco de dissertações /teses ainda não se encontra completo. Dos 420 documentos, o acervo está constituído por 205 dissertações/teses.

Do total de documentos - o objeto da pesquisa - foram analisadas 55 dissertações/teses e geradas duas listas por categoria semântica: uma dos termos existentes no **TESAURO...** (IBICT, 1989) e outra dos não existentes. Tem-se notado a necessidade de acrescentar novos termos ainda não incluídos no referido Tesauro.

As dificuldades encontradas nessa primeira etapa da pesquisa estão ligadas à imprecisão dos resumos das dissertações/teses.

Embora ainda inicial o trabalho, procurou-se: a) representar os assuntos dos documentos e das solicitações de busca; b) recuperar documentos por grandes áreas de assunto, com conteúdos semelhantes e/ou relevantes sobre um tema específico; c) auxiliar na escolha do termo adequado para a estratégia de busca e, finalmente, d) permitir a compatibilidade entre a linguagem do indexador e do usuário/pesquisador.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDREEWSKI, Alexandre & RUAS, Vitorino. Indexação automática em métodos linguísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informação**, Brasília, v.12, n.1, p.61-73, 1983.

CESARINO, Maria Augusta da Nobrega. Sistemas de recuperação da informação. **Rev.Esc.Bibliotecon.UFMG**, Belo Horizonte, v.14, n.2, p.157-68, set. 1985.

CINTRA, Ana Maria M. Elementos de linguística para estudos de indexação. **Ciência da Informação**, Brasília, v.12, n.1, p.5-22, 1983.

FUJITA, Mariângela Spotti Lopes. **PRECIS na língua portuguesa:** teoria e prática de indexação. Brasília: Editora Universidade de Brasília - ABDF, 1989. 213p.

GOMES, Hagar Espanha. O indexador face às novas tecnologias de informação. **Trans-in-formação**, Campinas, PUCCAMP, 1(2):161-71, maio/agosto, 1989.

GUTIERREZ, Antonio Garcia. Teoria de la indización: nuevos parámetros de investigación. **Trans-in-formação**, Campinas, PUCCAMP, 1(2):147-59, maio/agosto, 1989.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Diretrizes para elaboração de tesauros monolíngues**. Coordenado por Hagar Espanha Gomes. Brasília: IBICT, 1980. 70p.

_____. Manual de utilização de elaboração de tesauros em microcomputador (TECER). Brasília: IBICT, 1989. 60p.

KREMER, Jeanette M. Estratégia de busca. **Rev. Esc. Bibliotecon. UFMG**, Belo Horizonte, v.14, n.2, p.187-220, set. 1985.

LANCASTER, F.W. **Construção e uso de tesauro:** curso condensado. Trad. de César Almeida de Menezes Silva; Rev. de Odilon Pereira da Silva. Brasília: IBICT, 1987. 106p.

MANUAL de elaboração de tesauros monolíngues. Coordenado por Hagar Espanha Gomes. Brasília: Programa Nacional de Bibliotecas das Instituições de Ensino Superior, 1990. 78p.

NATALI, Johanna W. Documentação e linguística: inter-relação e campos de pesquisa. **Rev. Bras. Bibliotecon. Documentação**. [s.l.] v.11, n.11, n.1/2, p.33-42, jan/jun. 1978.

ROBREDO, Jaime. **Documentação de hoje e de amanhã**. Ed. do Autor, 1986. p.4-7.

_____. indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, 20(2), Brasília: 130-36, jul/dez. 1991.

TERMINOLOGY OF DOCUMENTATION: a selection of 1220 basic terms published in English, French, German, Russian and Spanish; compilet by gernot wersing and wrich neveling. Paris: The Unesco Press, 1976.

VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, 17(1), Brasília: 43-57, jan/jun. 1991.

Equipe de Pesquisa:

Allyson Vitale de Oliveira Lima
(Bolsa de Iniciação Científica)

Anea Lígia Benedito Duarte
(Bolsa de Aperfeiçoamento)

Cid Evangelista Júnior
(Bolsa de Aperfeiçoamento)

Iolanda Oliveira Rabaça
(Bolsa de Iniciação Científica)

Maurílio João Franchin Júnior
(Bolsa de Iniciação Científica)

Vânia Lando de Carvalho
(Mestranda)