INDEXAÇÃO AUTOMÁTICA E SEMÂNTICA:
estudo da análise do conteúdo de teses e dissertações

comunicação de trabalho

Graciane Silva Bruzinga Borges[*]
Benildes Coura Moreira dos Santos Maculan[**]
Gercina Ângela Borém de Oliveira Lima[***]

Resumo Objetiva avaliar a contribuição de técnicas específicas de indexação automática no processo de representação semântica do conteúdo de teses e dissertações. Descrevem-se os processos de Indexação Manual e de Indexação Automática e aborda-se a aplicação dos critérios sintático-semânticos na extração automática de termos relevantes para a representação do conteúdo de documentos acadêmicos. Discutem-se os referenciais teóricos advindos da semântica e da lingüística computacional. Para implementar o processo de indexação automática são apresentados o parser Tropes, para extração automática dos termos; e a Taxonomia da Ciência da Informação elaborada por Hawkins, Larson e Caton, em 2003, como cenário semântico embutido no software.

Palavras-chave INDEXAÇÃO AUTOMÁTICA
REPRESENTAÇÃO DA INFORMAÇÃO
SEMÂNTICA
SINTAXE
LINGÜÍSTICA COMPUTACIONAL

[*] Bibliotecária pela Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Mestranda em Ciência da Informação pela UFMG.
E-mail: gracianesb@yahoo.com.br

[**] Discente em Biblioteconomia pela Escola de Ciência da Informação da UFMG.
E-mail: benildes@gmail.com

[***] Professora Adjunta da Escola da Ciência da Informação da UFMG. Doutora em Ciência da Informação pela UFMG. Mestre em *Library and Information Science* pela Clark Atlanta University (EUA). E-mail: glima@eci.ufmg.br

1 Introdução

Este estudo apresenta um resultado parcial da pesquisa de mestrado[1] que visa avaliar a contribuição de técnicas específicas de indexação automática no processo de representação semântica do conteúdo de teses e dissertações em bibliotecas digitais. Acreditamos que esses documentos necessitam de tratamento perfeitamente adequado, por serem considerados objetos complexos, para que se obtenha uma recuperação das informações de forma mais eficiente. Esta pesquisa, por sua vez, é parte de projeto que objetiva a construção de um modelo hipertextual para organização de documentos acadêmicos e sua implementação tecnológica[2].

Indexar é o ato de selecionar ou definir termos (palavras ou expressões) que irão descrever o conteúdo de um determinado documento, sempre levando em consideração uma clientela específica. Assim, de acordo com Lancaster (2004), uma mesma publicação poderá apresentar conjuntos diferentes de termos de indexação dependendo do grupo de usuários

1 A pesquisa de mestrado referenciada encontra-se em desenvolvimento e é de autoria da primeira autora deste artigo, intitulada "Utilização de Técnicas de Indexação Automática para a Representação do Conteúdo Semântico de Documentos Acadêmicos". Pretende contribuir para elaboração do protótipo MHTX que irá tratar o conteúdo das teses e dissertações defendidas na Escola de Ciência da Informação da Universidade Federal de Minas Gerais, da linha de pesquisa Organização e Uso da Informação – OUI no ambiente de uma biblioteca digital.

2 O projeto referenciado é intitulado "Projeto de Pesquisa Modelagem Conceitual para Organização Hipertextual de Documentos Acadêmicos", que tem como objetivo a continuação da implementação do protótipo Mapa Hipertextual - MHTX. Esse protótipo teve implementação iniciada na pesquisa da tese de doutorado "Mapa Hipertextual (MHTX): um modelo para organização hipertextual de documentos" (LIMA, 2004). Esse projeto destina-se a servir às instâncias de: (1) produção primária (autores intelectuais) de documentos hipertextuais originais; (2) produção secundária (autores tecnológicos) na organização de sistemas hipertextuais ou bibliotecas digitais.

ao qual se destina e aos interesses particulares desse grupo, ou seja, não há uma forma 'correta' única de indexar. Segundo Navarro (1988), o ato de indexar define-se como traduzir o conteúdo de um documento em palavras que tornem possível sua recuperação. Entretanto, observam-se significativas dificuldades na interseção de lingüística e indexação, nesse processo.

> A capacidade íntima de reconhecer sobre o que trata o documento é a questão central do procedimento de indexação. [...]. Para fins de Indexação, o(s) termo(s) selecionado(s) é a correlação comportamental sobre o que se pensa 'sobre o que o documento trata', pois seria o termo usado para se procurar por tal documento (MARON, 1977 apud GUEDES, 1994).

Em síntese, indexar é substituir o texto de um documento por uma descrição abreviada de seu conteúdo, com o intuito de sinalizar sua essência. E essa representação é feita a partir da análise do conteúdo do texto-fonte, que, necessariamente, deveria ser feita por especialistas, sob o olhar atento de metodologias e procedimentos. Existem, pelo menos, duas formas para se fazer a análise de conteúdo semântico: indexação manual e indexação automática.

Na Ciência da Informação, os estudos sobre essas duas formas de indexação têm sido continuamente abordados pelos pesquisadores da área: GUEDES (1994); GIL LEIVA (1999); LANCASTER (2004); LOEHRLEIN, et. al. (2005); ROBREDO (1999; 2005); SILVA; FUJITA (2004). As investigações sobre a indexação manual apontam para alternativas que visam melhorar a abstração e tornar a representação da informação mais fidedigna à temática tratada pelo autor, cabendo ao indexador fazer esse processo manualmente. Em relação aos estudos sobre a indexação automática, nota-se que seu surgimento foi devido à necessidade de serem resolvidos problemas tais como a morosidade vivenciada na indexação manual, e como solução para agilizar os processos nos meios digitais.

Destaca-se aqui que a semântica é o meio utilizado para a representação do significado dos enunciados. As relações semânticas são importantes na estruturação do conhecimento e na formação de conceitos para escolha de termos representativos de significado.

Neste artigo são apresentados os critérios teóricos que tratam da importância da semântica e da estrutura sintática no processo de indexação automática e como o triângulo semântico de Ogden e Richards (1972), exposto na Teoria do Conceito de Dahlberg (1978), pode ser relacionado com esse contexto. Nesse trabalho, Dahlberg usa o triângulo semântico como modelo para a construção de conceitos e para representar as relações existentes entre o objeto, o conceito e o termo (LIMA, 2007).

Os critérios computacionais, tais como lógica de programação; algoritmos; fórmulas estatísticas e outros, imprescindíveis para o desenvolvimento de *softwares* de indexação automática, serão abordados em estudo posterior.

2 Indexação Manual e Indexação Automática

O processo manual de indexação pode ser divido em duas etapas essenciais: a análise conceitual e a tradução.

A etapa de análise conceitual determina do que trata um documento, isto é, qual seu assunto. Nessa atividade, a leitura e a compreensão do texto são primordiais, porém, o tempo restrito do indexador e a quantidade cada vez maior de documentos passíveis de tratamento são fatores preocupantes. "Ao indexador raramente é dado o luxo de poder ler um documento do começo ao fim" (LANCASTER, 1993, p. 20-21).

Para essa análise, é preciso considerar o domínio no qual o documento está inserido, identificando as características específicas do campo de conhecimento, sejam elas de ordem cultural, terminológica, históricas e lingüísticas. Para tanto, o conhecimento do indexador sobre esse domínio é importante para a qualidade dessa análise. Assim, a análise será feita contextualmente, pois o documento não será considerado como uma parte isolada, mas, como parte de um todo (HJORLAND, 1992).

A etapa de tradução objetiva converter o conteúdo do documento, determinado na etapa de análise conceitual, em um conjunto de termos de indexação; e essa transferência também é feita por meio de mediação semântica. Isso sempre acontecerá, mesmo em casos nos quais não houver prescrição de regras formais. Essas regras podem ser estipuladas em função dos interesses

da instituição ou do instrumento de controle terminológico. Esse controle é feito a partir do uso de termos autorizados retirados de algum vocabulário controlado, sendo que, muitas vezes, essa tarefa é feita de forma intuitiva. Alguns dos principais vocabulários controlados utilizados na Biblioteconomia são: Taxonomia; Lista de Cabeçalho de Assunto; Classificação Decimal de Dewey – CDD; e Classificação Decimal Universal – CDU.

Entretanto, a indexação manual vem revelando-se inadequada para minimizar a subjetividade inerente à indexação, além de ser caracterizada como um processo relativamente moroso e caro. Vários fatores podem ser apontados como causa deste problema. O conhecimento que o indexador tem sobre o assunto indexado determina o grau de consistência atingido. Tem-se, ainda, a dinamicidade do conhecimento, que exige do indexador permanente atualização. Outro aspecto a considerar, segundo Borko (1977 apud GUEDES, 1994), refere-se à inconsistência interindexadores (diferentes indexadores atribuindo diferentes termos-índice a um mesmo conceito/documento) e intraindexador (o mesmo indexador atribuindo diferentes termos-índice a um mesmo conceito/documento, em diferentes momentos). A possibilidade de o indexador não dominar o idioma do documento também é um fator que prejudica a qualidade da indexação.

Todos os problemas enumerados impulsionaram as pesquisas no campo da indexação automática, tornando-o bastante abordado pelos pesquisadores da Ciência da Informação. Alguns dos resultados dessas investigações apontam alternativas que podem trazer soluções interessantes para a área, sobretudo em ambientes digitais.

Segundo Robredo (1982), o processo de indexação automática é similar ao processo de leitura-memorização humano, sendo seu princípio geral baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecida, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas.

Ao ler um texto, não interessam, ao indivíduo, as letras, mas a idéia que elas representam quando organizadas em palavras ou em conjuntos de palavras. O olho, janela do cérebro, reconhece as palavras significativas e suas associações fixando-se nelas um tempo necessário para assegurar a memorização das idéias, pulando, praticamente, as palavras não significativas (ROBREDO, 1982).

Pode-se separar o processo de memorização humana em duas etapas principais: (1) memorização temporária e inconsciente, nessa etapa há a conservação das palavras significativas passando por uma modificação ou aperfeiçoamento das mesmas a partir da detecção de novos conceitos significativos; e (2) memorização permanente dos conceitos assim trabalhados, à qual se atribui o nome de memória. Depois de ocorridas as duas etapas, tem-se, no fim do processo, a fixação na memória de uma série de palavras-conceitos-descritores que representam as idéias básicas do documento que acabamos de ler. A leitura, através de um processo de análise-indexação, leva à armazenagem dos descritores que representam o conteúdo dos documentos (ROBREDO, 1982).

Pode-se destacar dois tipos de processos de indexação automática: (1) indexação por extração automática e (2) indexação por atribuição automática.

No Processo de Indexação por Extração Automática, palavras ou expressões que aparecem no texto são extraídas para representar seu conteúdo como um todo. Considerando uma versão eletrônica desse documento, é possível utilizar um programa computacional para extrair os termos a partir dos mesmos princípios utilizados por seres humanos, como: freqüência da palavra dentro do texto; posição da palavra no texto (no título, nas legendas, no resumo etc.) e por seu próprio contexto (LANCASTER, 2004). Na década de 1950 teve início a indexação automática baseada em freqüência com os trabalhos de Luhn, em 1957, e de Baxendale, em 1958. Baxendale (1958 apud LANCASTER, 2004) sugere que, em substituição ao processo que analisa todo o texto, sejam analisadas apenas o "Tópico Frasal" e as "Palavras Sugestivas". Seus estudos demonstraram que era necessário o processamento apenas da primeira e da última frase de cada parágrafo, pois, em 85% das vezes a primeira frase era o tópico frasal e em 7% dos casos a última frase o era. Considera-se como tópico frasal a parte do texto que provia o máximo de informações relativas ao conteúdo do texto. Os sistemas baseados em

indexação por extração automática realizam, basicamente, as seguintes tarefas: (1) contar palavras num texto; (2) cotejá-las com uma lista de palavras proibidas; (3) eliminar palavras não significativas (artigos, preposições, conjunções, etc.); (4) ordenar as palavras de acordo com sua freqüência.

O Processo de Indexação por Atribuição Automática é mais complexo de ser realizado com maior eficiência que o processo de indexação por extração automática. Em geral, é considerada uma atividade difícil, pois, para a representação do conteúdo temático, é necessário um controle terminológico. Deve-se desenvolver, para cada termo atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer nos documentos. Por exemplo, para o termo 'chuva ácida' incluir-se-iam as expressões 'precipitação ácida', 'poluição atmosférica', 'dióxido de enxofre' etc. Um problema relevante nesse processo pode ser ilustrado com a seguinte situação: a frase "dois dias depois de a substância haver sido ingerida surgiram diversos sintomas" pode ser legitimamente indexada por uma pessoa sob o assunto 'toxidade'. Já para um *software*, essa tarefa é verdadeiramente difícil (O'CONNOR, 1965 apud LANCASTER, 2004). Esse tipo de indexação automática remonta a uma longa história. Tentativas iniciais não obtiveram muito êxito, porém, nos últimos 40 anos têm-se resultados melhores nessa área (BORKO; BERNICK, 1963 apud LANCASTER, 2004).

O histórico da indexação automática pode ser associado com o uso de programas computacionais para geração de índices pré-coordenados. Segundo Lancaster (2004, p.50), "a flexibilidade inerente aos sistemas pós-coordenados deixa de existir quando os termos de indexação são impressos em papel ou fichas catalográficas convencionais". Dois exemplos de índices pré-coordenados são os índices impressos e os catálogos, ambos caracterizados por: (1) dificuldade de representação de multidimensionalidade das relações entre os termos; (2) possibilidade de listagem dos termos somente em uma seqüência, o que implica que o primeiro termo é mais importante que os demais; (3) dificuldade, ou impossibilidade, de combinação de termos no momento da busca (LANCASTER, 2004, p.50).

Segundo Lancaster (2004, p. 52), "vários programas de computador foram desenvolvidos para gerar, automaticamente, um conjunto de entradas de índices a partir de uma seqüência de termos". Como exemplos podem-se citar os modelos SLIC, o PRECIS, o KWIC, o KWOC e o NEPHIS, descritos a seguir.

O *Selective Listing in Combination* – SLIC (Listagem Seletiva em Combinação) – foi criado por J. R. Sharp em 1966. O programa organiza a seqüência de termos de um documento em ordem alfabética e elimina as seqüências redundantes. Já o método PRECIS produz índice impresso baseado na ordem alfabética e na 'alteração' sistemática de termos para que ocupem a posição de entrada. Modelos como o SLIC pressupõem o emprego de termos de indexação e não de texto livre[3].

Entretanto, desenvolveram-se métodos bem mais simples para a construção de índices a partir de texto, especialmente a partir de palavras que ocorrem nos títulos dos documentos. São exemplos desses métodos o *Keyword in Context* – KWIC (Palavra-chave no Contexto) – e o *Keyword out of Context* – KWOC (Palavra-chave fora do Contexto) – (LANCASTER, 2004, p.54).

Para Robredo (1982, p. 238), a primeira aplicação generalizada da indexação automática de documentos técnicos, a partir de palavras significativas dos títulos, se deu com

O KWIC foi desenvolvido por H. P. Luhn em 1959 e corresponde a um índice rotativo em que cada palavra-chave que aparece nos títulos dos documentos torna-se uma entrada do índice. Cada palavra-chave é destacada de alguma forma e as palavras restantes do título aparecem 'envolvendo-a'. O critério usado para selecionar as palavras que irão compor o índice é chamado de processo 'reverso', ou seja, o programa reconhece as palavras que não são palavras-chave, baseado em uma lista de palavras proibidas, e impede que elas sejam adotadas na entrada. Os vocábulos dessa lista de palavras proibidas têm função sintática (artigos, preposições, conjunções, etc.), mas em si mesmos não representam conteúdo temático. O KWIC é um método simples, barato e que obtém, em certo nível, acesso temático ao conteúdo de uma coleção. É nítido, porém, que sua qualidade está diretamente relacionada à qualidade dos títulos, no sentido em que estes

3 De acordo com Lancaster (2004), no contexto da recuperação da informação, texto livre corresponde às palavras que ocorrem em textos impressos, podendo ser o título, um resumo, um extrato, ou o texto integral de uma publicação.

devem ser bons indicadores do conteúdo dos textos (LANCASTER, 2004, p.54, 55).

O método KWOC é semelhante ao KWIC, porém as palavras-chave que se tornam pontos de acesso são repetidas fora do contexto, normalmente destacadas no canto esquerdo da página ou usadas com cabeçalhos de assunto.

Ainda segundo Lancaster (2004, p.56-59), o *Nested Phrase Indexing System* – NEPHIS (Sistema de Indexação de Frase Encaixada) – é um índice articulado de assunto e foi criado por T. C. Craven em 1977. Índice este que foi descrito minuciosamente por Armstrong e Keen em 1982. Nesse modelo, os termos de entrada são reordenados de tal modo que cada um deles se liga a seu vizinho original por meio de uma palavra funcional ou pontuação especial, conservando-se, assim, a estrutura similar à de uma frase, mesmo que muitas vezes disposta em ordem diferente.

Visualizando o modelo descrito anteriormente, fica claro que a sintaxe do texto original é mantida, de modo que o significado do enunciado não fica obscuro. Nesse sentido, o processo de indexação automática deve respeitar e incluir os princípios da semântica e da sintaxe.

3 O Papel da Semântica e da Sintaxe na Indexação Automática

É sabido que toda língua tem seu próprio recorte e sua própria semântica, que, como dito anteriormente, estuda os significados das coisas. Essa língua pode ser repleta de regionalismos, metáforas, gírias, linguagem figurada, denotação e conotação. Tudo aquilo que está presente na vida das pessoas possui um nome, que é parte do léxico. A estrutura lexical compreende o conjunto de vocábulos de uma língua e abrange o conhecimento lingüístico partilhado pela sociedade na qual é falada, possuindo valor diferente de língua para língua. A análise sintática consegue determinar se uma expressão ou frase está adequada à gramática dessa língua específica. Temos ainda, a unidade lexicalizada, que pode conter várias palavras com significado convencionado, como, por exemplo, *bater as botas* (morrer) ou *dar com os burros n'água* (fracassar). Nesse contexto, quando entendemos que a expressão *bater as botas* é morrer, o que fazemos é dar um novo valor semântico à expressão.

Para Rector e Yunes (1980, p. 14),

> uma explicação de propriedades semânticas requer mais do que a análise do sentido das palavras apenas, isto é, para que se entenda o sentido de uma sentença e suas relações semânticas com outras expressões, é preciso saber não só o significado de suas unidades léxicas, mas, também, como estas se relacionam – a dependência da estrutura sintática da sentença.

Por exemplo, retomando a expressão bater as botas (FIG. 1):

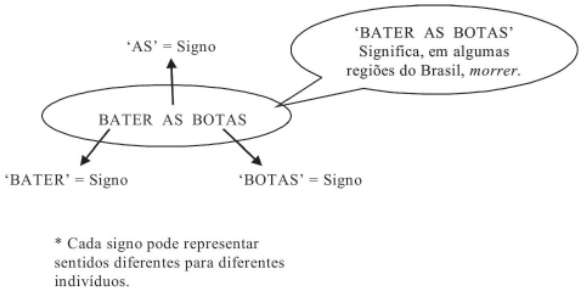

Figura 1: A semântica de uma frase
Fonte: Desenvolvida por Borges; Maculan; Lima (2007) para a pesquisa em questão.

O processo de análise conceitual também demanda forte carga cognitiva, efetuado pela mente humana, que é a interpretação, a coordenação de signos e a abstração de significados. Signo é uma palavra que, isoladamente, pode ter sentido para determinado indivíduo, mas não possui significado. O significado de um signo (uma palavra) está estreitamente ligado à estrutura lexical, isto é, ao contexto no qual o signo está inserido. Já o sentido é uma abstração pessoal, ou seja, é como cada indivíduo entende o signo. Sobre isso, temos a afirmação:

> Considerado isoladamente, signo algum tem significação [significado]. Toda significação de signo nasce de um contexto, quer entendamos por isso um contexto de situação ou um contexto explícito [...]. É necessário, assim, abster-se de acreditar que um substantivo está mais carregado de sentido do que uma preposição, ou que uma palavra está mais carregada de significação do que um sufixo de derivação ou uma terminação flexional (HJELMSLEV, 1975, p.50 apud SILVA, [2004]).

A semântica e a sintaxe têm papéis importantes na indexação automática, na medida em que permitem ao *software* identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento.

A sintaxe determina a forma correta de construção das frases de uma determinada língua, levando em consideração a seqüência de sujeitos, verbos, objetos, predicados, artigos, preposições etc. A semântica, por sua vez, se encarrega do significado da frase construída. Dessa forma, podemos ter frases sintaticamente corretas, mas sem um conteúdo semântico denotativo aceitável, e vice-versa. Exemplos:

- "A chuva gosta de cair sobre meus cabelos ruivos" = Frase com sintaxe correta, porém, sem conteúdo semântico denotativo aceitável.
- "Fingimos que fumus e vortemos" = Frase semanticamente inteligível, porém, fora dos parâmetros da língua portuguesa formal.

A sintaxe permite apenas escrever frases corretas numa língua. Por exemplo, as frases "O rato come o queijo" e "O queijo come o rato" são sintaticamente corretas, porém, assumimos que apenas a primeira frase tem significado na nossa língua. Isso se deve ao nosso conhecimento de que ratos são animais que se alimentam de queijo e de que é impossível, dentro da realidade, um queijo comer um rato.

Durante a análise sintática, pode-se perceber se os sintagmas foram colocados na seqüência correta. Sintagmas são expressões que ditam uma relação de dependência, na qual um elo de subordinação é estabelecido e cada um dos elementos é também um sintagma. Esse termo é geralmente empregado para designar cada parte de uma oração e pode ser: sintagma nominal (nome); sintagma adjetival (adjetivo); sintagma verbal (verbo); sintagma preposicional (preposição); e sintagma adverbial (advérbio). Conseguir identificar os sintagmas é muito importante na análise sintática, porque isso facilitará a compreensão do papel exercido pelas palavras na frase. Como ilustração, temos a frase: O *Christiano acreditou na vitória*, analisada na FIG. 2.

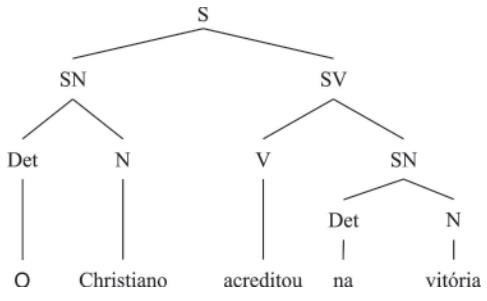

Figura 2: Árvore 2sintagmática
Legenda: S = sentença (frase)
SN = sintagma nominal
Det = determinante
N = nome ou substantivo
SV = sintagma verbal
V = verbo
Fonte: Adaptado de Othero e Menuzzi (2005, p. 49).

Para determinar um sintagma é necessário identificar o elemento núcleo, uma vez que ele pode possuir mais de uma palavra. Ainda pode existir, numa mesma frase, mais de um sintagma de mesmo tipo, como é o caso do exemplo acima: *Christiano* e *vitória* são núcleos dos sintagmas nominais. Nesses casos, é preciso estabelecer quais funções sintáticas esses núcleos desempenham. No caso de nosso exemplo, seria: *Christiano* tem função de sujeito e *vitória* tem a função de objeto indireto.

Em síntese, para que um *software* de indexação automática apresente resultados satisfatórios, entre vários outros critérios possíveis e importantes, é essencial que ele seja capaz de analisar um texto tanto sob seu aspecto sintático quanto no semântico.

Um forte aliado na construção de um *software* com essas características de análise é a Lingüística Computacional, uma vez que essa área de pesquisa engloba as áreas da Informática e da Lingüística. A Lingüística Computacional tenta encontrar soluções tecnológicas para que as máquinas possam reproduzir o conhecimento lingüístico e semântico de um ser humano.

4 A Lingüística Computacional

A área de Lingüística Computacional preocupa-se com a compreensão da língua e de técnicas apropriadas à interpretação dessa língua, escrita ou falada, tentando imitar a capacidade

humana de comunicação. Para tanto, essa área utiliza elementos de sintaxe, semântica, fonética e fonologia, pragmática e análise do discurso. De acordo com Othero e Menuzzi (2005), a Lingüística Computacional pode ser dividida em Lingüística de Corpus e Processamento da Língua Natural (PLN).

Ainda segundo esses autores, a Lingüística de Corpus trabalha com o "*corpora* eletrônicos", isto é, grandes bancos de dados que contenham amostras de linguagem natural, que podem ser de diferentes fontes. O objetivo é estudar os fenômenos lingüísticos que podem acontecer em grandes amostras de uma língua específica e não, necessariamente, produzir um *software*. Esses *corpora* podem ser compostos de linguagem falada, ou de amostras de textos de várias fontes escritas, como por exemplo, textos jornalísticos. Já a PLN, preocupa-se com o estudo da linguagem diretamente voltado para a construção de *softwares*, como tradutores automáticos, *chatterbots*, reconhecedores automáticos de voz, geradores automáticos de resumos, *parsers,* entre outros.

Um *parser*,

> [...] no contexto da lingüística computacional, é um analisador automático (ou semi-automático) de sentenças [frases]. Esse tipo de programa é capaz de analisar uma sentença com base em uma gramática preestabelecida de determinada língua, verificando se as sentenças fazem parte ou não da língua, de acordo com o que autoriza a sua gramática. Um *parser* também analisa sintaticamente as sentenças, decompondo-as em uma série de unidades menores, primeiramente em nódulos não-terminais (os sintagmas), até chegar a nódulos terminais (os itens lexicais), atribuindo-lhes uma estrutura de constituintes. Essa estrutura de constituintes, que representa a representação hierárquica e sintática da frase, é apresentada comumente em árvores sintáticas ou através de colchetes rotulados (OTHERO; MENUZZI, 2005, p. 49).

Um *parser* utiliza linguagem do tipo "declarativa", essa linguagem fornece proposições ao computador com as quais ele é capaz de analisar as frases de um texto por meio de combinações lógicas. A maioria das linguagens de programação geralmente é do tipo "procedural", que, ao contrário da linguagem do tipo "declarativa", provê o computador de um algoritmo cujos passos são ações executadas pelo próprio computador até que ele chegue a determinado resultado (OTHERO; MENUZZI, 2005, p. 42).

Ainda na abordagem da lingüística computacional, gramática é entendida como um conjunto relativamente pequeno de regras e vocábulos de uma língua, que possibilita reconhecer todas as frases possíveis dessa determinada língua, atribuindo a elas uma estrutura sintagmática; essa gramática é denominada 'gramática sintagmática'. Já no contexto da Lingüística, o termo *gramática* é entendido sob diferentes acepções, desde as normas que regem uma língua até o sentido de gramática histórica e de gramática comparada (OTHERO; MENUZZI, 2005).

Nota-se, assim, que o tratamento computacional para o processamento semântico das línguas naturais ainda se encontra em estágios iniciais de pesquisa para que possam trazer as soluções almejadas. Uma das soluções apontadas na literatura para reduzir os problemas de processamento da linguagem natural é o uso de linguagens controladas, nas quais existe uma estrutura semântica coesa, com relações terminológicas preestabelecidas, dentro de um determinado domínio.

6 Vocabulário Controlado: Taxonomia

Para o tratamento da produção científica pode-se utilizar a linguagem natural ou a linguagem controlada. Na linguagem natural, o termo (palavra ou expressão) será retirado do texto. Na linguagem controlada, há uma lista de termos escolhidos, cuja função é a de só admitir uma forma de interpretação, ou seja, de significado, além de possibilitar uma maior padronização e rigor de utilização de termos. Nesse contexto, a Biblioteconomia estuda as possibilidades de elaboração de linguagens documentais que possibilitem identificar o conteúdo, isto é, termos (palavras e expressões) mais significativos e estabelecer relações semânticas entre esses termos, por meio de hierarquias. O principal objetivo dessas investigações é facilitar a representação temática do conteúdo de um documento e indexá-lo.

Dessa forma, a taxonomia, sendo um instrumento de controle terminológico, torna-se uma ferramenta importante na representação semântica de documento. Como ferramenta especializada, é construída por meio de um processo que visa arranjar hierarquicamente uma lista de conceitos que representam a temática de determinado domínio ou área. As taxonomias devem atender a diversos tipos de objetivos e podem ser apresentadas na forma de representações gráficas, facilitando a compreensão e exploração do conteúdo (FIGUEIREDO, 2006).

Uma taxonomia,

> [...] em linhas gerais, é a área do conhecimento que se ocupa das regras e dos princípios da nomenclatura. Pode ser vista como um sistema de classificação tendo por base, normalmente, uma hierarquia de termos e conceitos, na qual os termos localizados nos níveis mais baixos representam os aspectos mais específicos do conteúdo. Até recentemente, o seu interesse era restrito a profissionais da área de ciência da informação, biblioteconomia ou especialistas em determinadas ciências, como a biologia, mas agora é parte do interesse dos profissionais da gestão do conhecimento. A correta definição e classificação das bases de conhecimento de uma empresa, ou seja, uma estrutura adequada de termos e conceitos tornou-se fundamental para a gestão da Intranet, portais etc. (GLOSSÁRIO NETIC, [200-]).

Quando uma taxonomia assume interface gráfica, as informações dispersas no texto são organizadas, respeitando-se os temas, assuntos e a hierarquia estipulados pela ferramenta. Dessa forma, será extraído o que há de mais relevante naquele contexto.

"A taxonomia define classes, subclasses e as relações entre elas, e o conjunto de regras de inferência fornece o mecanismo de manipulação dos objetos das classes, utilizando raciocínio lógico" (PICKLER, 2007, p. 73).

Essas características de uma taxonomia tornam-na um instrumento de importância fundamental no processo da indexação automática, pois ela permite a representação do conteúdo. O conteúdo de um documento é representado por uma taxonomia a partir de termos e definições utilizados num domínio específico. Baseado nessas premissas, partimos para a construção de uma taxonomia como uma ferramenta auxiliar na proposta de indexação automática, descrita a seguir, na tentativa de conciliar a intervenção humana a um processo automático.

6 Proposta de indexação automática utilizando critérios sintático-semânticos

O mercado oferece *softwares* de indexação automática que prometem realizar a extração de termos relevantes para a representação do conteúdo informacional, baseados em critérios semelhantes aos utilizados pelos seres humanos.

Nesta seção apresentaremos um *parser* de extração automática que utiliza os critérios sintático-semânticos. Esta exposição será restrita a uma avaliação teórica da importância do uso desses critérios, tendo em vista que os testes necessários para a verificação da eficácia, ou não, desse *parser*, serão realizados em uma etapa posterior da pesquisa. Além disso, será apresentada uma comparação do triângulo semântico, exposto na Teoria do Conceito de Dahlberg, com a utilização da taxonomia utilizada como cenário para enriquecer e filtrar o conteúdo semântico das teses e dissertações.

Para o desenvolvimento desta etapa do trabalho de pesquisa, serão utilizadas as seguintes ferramentas: o *parser* Tropes e a taxonomia da área de Ciência da Informação construída por Hawkins, Larson e Caton, elaborada em 2003, como cenário semântico. Ambos serão descritos a seguir.

O *parser* Tropes foi criado em 1994 e, em sua primeira versão, era capaz de analisar obras literárias do tipo romance. Atualmente, ele se tornou um motor semântico, funcionando em seis línguas, entre elas o português de Portugal e o português do Brasil.

Para analisar o conteúdo utilizando critérios sintático-semânticos, o Tropes usa os recursos de uma gramática sintagmática e de um cenário padronizado determinado previamente para a análise. Essa gramática já vem embutida no Tropes, abrangendo substantivos, verbos, adjetivos, determinantes, conectores, modalizações e pronomes relativos e pessoais. Já o cenário precisa representar o conteúdo semântico da área a ser analisada. Ele possui a vantagem de permitir que o utilizador

construa seu próprio cenário, adequando-o aos seus objetivos. Portanto, nesta nossa proposta, construiremos um cenário a partir de uma taxonomia da área de Ciência da Informação.

Na análise lexico-semântica, o Tropes irá detectar as palavras que representem o conteúdo de uma frase, agrupando-as em classes de equivalentes determinadas pelo cenário. Se o cenário apresentar uma palavra em determinada classe, o Tropes será capaz de resolver problemas de ambigüidade. Assim, se nesse cenário a palavra <cultura> estiver na classe <Agricultura>, ela sempre será associada a essa classe, independentemente do texto utilizado. Além disso, o Tropes calcula a probabilidade da ocorrência da palavra <cultura> nessa classe, com uma taxa média de erro de 5%. Outro exemplo de aplicação desse recurso é a capacidade do Tropes de fazer a distinção entre o rato (animal) do rato (*mouse* do computador), a partir de cenários que poderiam tratar, por exemplo, de <Veterinária> e <Informática>, respectivamente.

O Tropes consegue fazer, ainda, uma análise morfo-sintática. Nessa análise, a categoria morfológica da frase é identificada, por meio da análise da origem, da formação e do significado de cada uma das palavras, além de sua flexão (de gênero, grau e número). Na análise morfológica, por exemplo, a frase *O Tropes é um software* seria destrinchada: *O* = artigo definido; *Tropes* = substantivo próprio; *é* = verbo; *um* = artigo indefinido e *software* = substantivo comum.

Para a utilização do Tropes neste trabalho de pesquisa, ao invés de criar o próprio cenário semântico baseado na terminologia da linha de pesquisa Organização e Uso da Informação – OUI – optou-se por escolher um cenário mais abrangente, porém, que possibilite ao *software* trabalhar contextualmente nesse domínio. Assim, esse será um *cenário especializado*, que poderá ser mudado constantemente, com a inserção de novos termos facilitando a filtragem de termos equivalentes, garantindo assim a consistência na indexação.

Um dos instrumentos que podem ser utilizados como cenário especializado é uma taxonomia, que pode ser definida, de forma literal, como a organização de um conjunto de termos a partir de princípios estabelecidos. O princípio mais utilizado é o de uma taxonomia hierárquica, construída em níveis de generalidades.

Como mencionado anteriormente, para a construção desse nosso *cenário especializado*, foi utilizada a Taxonomia da Área de Ciência da Informação, elaborada por Hawkins, Larson, e Caton, e proposta no *Information Science Abstracts*, cuja tradução está incorporada ao trabalho das autoras Oddone e Gomes (2003). Nesse trabalho, também é apresentada uma taxonomia da mesma área, de autoria própria. A taxonomia escolhida apresenta objetivos muito similares aos objetivos da taxonomia de Oddone e Gomes (2003), apesar de haver algumas diferenças, principalmente em relação à ênfase, à nomenclatura, e à organização e detalhamento.

Cabe justificar, aqui, que a decisão pela escolha da Taxonomia de Hawkins, Larson e Caton foi baseada unicamente no fato de que este instrumento apresenta categorias mais específicas e que estas contemplam mais detalhadamente as temáticas tratadas na área de organização e uso da informação.

A utilização da taxonomia como cenário especializado no Tropes nos remete à análise conceitual feita partindo do princípio do triângulo semântico, apresentado no trabalho de Dahlberg (1978).

De acordo com o proposto por Dahlberg (1978), um indivíduo, ao olhar um referente (objeto), irá, a partir de seus conhecimentos prévios, memória e experiências anteriores, atribuir características (conceitos) a esse referente. A partir dessas características, ele irá atribuir uma forma verbal[4] (termo) que represente ou sintetize plenamente e convenientemente esse referente. Cada indivíduo poderá atribuir formas verbais diferentes para um mesmo referente. Essa maneira de identificar significados pode ser adequada na vida cotidiana, mas quando se trata de indexação de documentos científicos, isso não seria muito apropriado.

As características (conceitos) podem ser consideradas como a matéria-prima do processo de indexação. Para Dahlberg (1978), 'conceito' é a análise e a síntese de enunciados verdadeiros sobre um objeto, e pode ser entendido como uma "unidade de conhecimento", representado por uma forma verbal.

Para Campos e Gomes (2006), o "conceito é, de fato, o ponto de partida para estabelecer as relações conceituais e determinar a forma verbal mais adequada para representá-lo". Temos assim, o triângulo de Dahlberg (FIG.3):

4 No contexto do trabalho da Dahlberg a forma verbal possui um sentido mais amplo e refere-se a palavra ou expressão, e não se limita apenas ao sentido de classe verbal.

Figura 3: Triângulo de Dahlberg
Fonte:Aadaptado de Dahlberg (1978).

Para Lima (2007, p. 2),

> O processo mental da formação do conceito se dá através de uma linha de pensamento que leva à elaboração do conhecimento, passando por um processo de assimilação da informação pelo cérebro, transformando-a. Após essa elaboração mental, baseada no conhecimento prévio do indivíduo, a unidade de informação se transforma em uma unidade conceitual que é representada por um termo [forma verbal], o qual possui um único significado, geralmente expresso por símbolos e palavras, com o objetivo de comunicação. Conceitos e categorias são temas centrais de estudo dentro da área de ciência da informação, sendo a base para a organização e para a representação do conhecimento.

O emprego da taxonomia de Hawkins, Larson e Caton, como cenário especializado para auxiliar a filtragem da terminologia de cada tese, pode ser transposto para os princípios do triângulo de Dahlberg através de dois outros triângulos, exemplificados a seguir (FIG. 4).

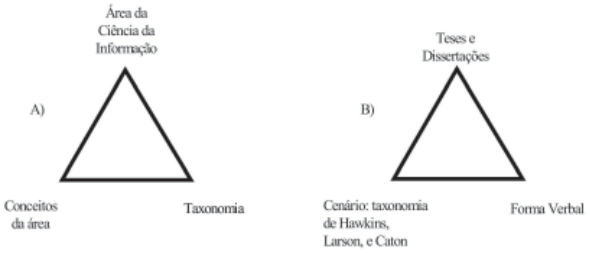

Figura 4: Relação do triângulo de Dahlberg com a construção e uso de taxonomias
Legenda:
a) Construção de uma taxonomia da área de Ciência da Informação
b) Aplicação da taxonomia de Hawkins, Larson e Caton na representação de teses e dissertações
Fonte: Desenvolvida por Borges; Maculan; Lima (2007) para a pesquisa em questão.

Fazendo relação com a FIG. 3, o triângulo A representa, primeiramente, como *Referente* a *Área da Ciência da Informação*, representada pelo conteúdo temático disponível na literatura especializada na área. As *Características* são os *Conceitos da área* desse domínio, identificados por meio de uma análise conceitual. A *Forma Verbal* é a *Taxonomia,* desenvolvida para representar mais adequadamente os conceitos.

O triângulo B representa a aplicação do critério semântico transportado para o processo de indexação automática proposto. Nele, o *Referente* será representado pelas *Teses e Dissertações*, selecionadas da linha OUI, que serão inseridas no *parser* Tropes; e as *Características* são representadas pelo *Cenário: Taxonomia da Ciência da Informação de Hawkins, Larson e Caton* a ser utilizada como um ambiente conceitual validado. A *Forma Verbal* será o resultado final dessa indexação automática, ou seja, o conjunto de termos que representará o conteúdo das teses e dissertações.

Finalmente, esse conjunto de termos extraídos será analisado e organizado sob critérios ainda não estipulados, permitindo a avaliação da eficácia, ou não, das ferramentas utilizadas nesse processo.

7 Considerações finais

Este trabalho nos está permitindo investigar o processo de indexação automática e as teorias nas quais ele se baseia. Embora ainda não possamos relatar os resultados finais desta pesquisa, podemos vislumbrar algumas considerações.

A indexação é o elo forte entre o que é disponibilizado no sistema e a necessidade do usuário. A fase de análise de conteúdo é a mais importante e, em contrapartida, a mais morosa para o indexador, principalmente quando este quer fazer um trabalho bem feito. A atividade de indexar tem-se tornado cada vez mais intensa, desde que as publicações se multiplicaram. As constantes buscas por informação pelo usuário propiciam um cenário no qual se faz necessário organizar as informações de forma sistemática, para disponibilizá-las.

Concluímos, a partir das investigações realizadas e nas atividades práticas laboratoriais com o *parser* Tropes, que para que a parte conceitual do protótipo MHTX seja consistente e

eficiente, exige a automatização do processo de indexação dos documentos acadêmicos que irão compor a base da biblioteca digital proposta. Isso deverá ser feito a partir de critérios sintáticos e semânticos de um documento no processo de indexação.

São muitos os estudos que atentam para a solução desse problema e há *softwares* construídos com diferentes modelos de indexação automática. Todos visam otimizar a atividade de análise de conteúdo, uma vez que conseguem fazer uma leitura quase instantânea do texto, muitas vezes com mais coerência e menos tendenciosidade que um ser humano. As pesquisas na área da Linguagem Computacional estão vislumbrando soluções que irão possibilitar às máquinas reproduzirem o conhecimento lingüístico e semântico de um ser humano.

Além disso, acreditamos que a adoção de uma taxonomia como cenário semântico usado no *parser* de teste é essencial para um resultado satisfatório. A taxonomia de Hawkins, Larson e Caton, escolhida para tal tarefa se mostrou consistente, pois apresenta termos pertinentes na cobertura da área, de forma geral, e termos atuais quanto à subárea computacional.

A relação estabelecida entre o triângulo de Dahlberg (1978) e o uso da taxonomia está sendo usada como principio norteador na tentativa de uma aproximação da indexação automática pretendida pelo protótipo MHTX ao processo cognitivo realizado pelo ser humano nesta atividade.

O que se espera na atualidade, e que deve ser uma das prioridades da área, é que esses instrumentos de indexação automática sejam ferramentas capazes de, concomitantemente, minimizar a subjetividade do indexador e imitar o raciocínio humano. Eles deverão levar em consideração o contexto semântico, respeitando princípios teóricos consistentes.

Esperamos que, nesta pesquisa, o processo de indexação automática venha suprir os problemas inerentes à indexação manual descritos, agilizando a seleção dos termos dos documentos acadêmicos do Programa de Pós-graduação em Ciência da Informação da ECI/UFMG. Isso será uma importante contribuição para a efetiva implementação do protótipo Mapa Hipertextual - MHTX.

SEMANTIC AND AUTOMATIC INDEXING: study of subject analysis of thesis and dissertations

ABSTRACT *This study aims at evaluating the specific contributions of automatic indexation techniques for the semantic representation process of dissertations' and thesis' contents. It describes the manual and automatic indexation processes and approaches the application of the semantic-syntactical criteria used in the automatic extraction of relevant terms in the representation of academic documents. It discusses theoretical references drawn from the Semantics and Computational Linguistics. To implement the automatic indexation process the parser Tropes for automatic extraction of the terms and the Taxonomy of Information Science are presented (Hawkins, Larson and Caton, 2003) as a semantic context built in the software.*

Keywords *AUTOMATIC INDEXATION*
REPRESENTATION OF THE INFORMATION
SEMANTICS
SYNTAX
COMPUTATIONAL LINGUISTIC

Artigo recebido em 13.12.2007 e aceito para publicação em 28.02.2008

Referências

BARQUIN, Beatriz A. R.; GONZÁLEZ, José A. M.; PINTO, Adilson L. Construção de uma ontologia para sistemas de informação empresarial para a área de telecomunicações. **DataGramaZero - Revista Ciência da Informação**, Brasília, DF, v. 7, n. 2, abr. 2006.

BORKO, H. Toward a theory of indexing. Information Processing and Management, v. 13, p. 355-365, 1977 apud GUEDES, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da Informação**, Brasília, DF, v. 23, n. 3, p.318-326, set./dez. 1994.

CAMPOS, Maria Luiza Almeida; GOMES, Hagar Espanha. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 11, n. 3, set./dez. 2006. Disponível em: <http://www.scielo.br/scielo.php? script=sci_arttext&pid=S1413-9362006000300005&lng=enenandothers&nrm=iso&tlng= enenandothers>. Acesso em: 21 jul. 2007.

DAHLBERG, Ingetraut. Teoria do conceito. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 2, p. 101-107, jul./ dez. 1978.

FIGUEIREDO, Saulo. **O impacto da taxonomia nas empresas.** [S.l.]: Webinsider, 28 nov. 2006. Disponível em: <http://webinsider.uol.com.br/index.php/2006/11/28/a-importancia-e-o-impacto-da-taxonomia-nas-empresas/>. Acesso em: 24 jul. 2007.

GIL LEIVA, Isidoro. **La automatización de la indización de documentos.** Madrid: Ediciones Trea, 1999. 221 p.

GUEDES, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da Informação**, Brasília, DF, v. 23, n. 3, p. 318-326, set./dez. 1994.

HAWKINS, Donald T.; LARSON, Signe E.; CATON, Bari Q. Information science abstracts: tracking the literature of information science. Part 2: a new taxonomy for information science. **Journal of the American Society for Information Science and Technology,** v. 54, n. 8, p. 771-781, 2003.

HJELMSLEV, Louis. Prolegômenos a uma teoria da linguagem. Tradução de: J. Teixeira Coelho Netto. São Paulo: Perspectiva, 1975 apud SILVA, Antônio Carlos da. **As teorias do signo e as significações lingüísticas**. [2004]. [Texto online]. Disponível em: <http://www.partes.com.br/ed39/teoriasignosreflexaoed39.htm>. Acesso em: 10 jul. 2007.

HJORLAND, Birger. The concept of 'subject' in Information Science. **Journal of Documentation**, v. 48, n. 2, p.172-200, June 1992.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004. 452 p.

____. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 1993. 347p.

LEROY, M. **As grandes correntes da lingüística moderna.** Trad. de Izidoro Blikstein e José Paulo Paes. São Paulo: Cultrix, 1971. 194 p.

LIMA, G. A. B. *Categorização como um processo cognitivo*. **Ciências & Cognição**, ano 4, v. 11, p.156-167, 2007. Disponível em: <www.cienciasecognicao.org>. Acesso em: 9 ago. 2007.

LOEHRLEIN, Aaron, et. al. A hybrid approach to faceted classification based on analysis of descriptor suffixes. In Grove, Andrew, Eds. Proceedings 68th **Annual Meeting of the American Society for Information Science and Technology (ASIST)** 42, Charlotte (US), p. 1-25. 2005.

LUHN, H.P. A statistical approach to mechanized encoding and searching of library information. **IBM J. Res. Dev.**, v. 1, n. 4, p. 309-317, 1957.

NAVARRO, Sandrelei. Interface entre lingüística e indexação: uma revisão de literatura. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n. 1/2, p. 46-62, jan./jun. 1988.

ODONNE, Nanci; GOMES, Maria T.F.S. Os temas de pesquisa em ciência da informação e

suas implicações político-epistemológicas. In: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO: CINFORM, 5., Salvador, 2004. **Anais...** Salvador: UFBA, 2004. Disponível em: <http://www.cinform.ufba.br/v_anais/artigos/nancioddone.html>. Acesso em: 2 jul. 2007.

OGDEN, C. K.; RICHARDS, I. A. **O significado de significado:** *um estudo da influência da linguagem sobre o pensamento e sobre a Ciência do Simbolismo*. Rio de Janeiro: Zahar, 1972. 348 p.

OTHERO, Gabriel de Ávila; MENUZZI. Sérgio de Moura. **Lingüística computacional:** teoria e prática. São Paulo: Parábola, 2005. 126 p.

PICKLER, Maria Elisa Valentim. Web semântica: ontologias como ferramentas de representação do conhecimento. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 65-83, jan./abr. 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100006&lng=en&nrm=iso&tlng=en>. Acesso em: 24 jul. 2007.

RECTOR, Monica; YUNES, Eliana. **Manual de semântica.** Rio de Janeiro: Ao Livro Técnico, 1980. 171 p.

ROBREDO, Jaime. **Documentação de hoje e de amanhã:** uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília: Reproart, 2005. 409 p.

ROBREDO, Jaime. Indexação e recuperação da informação na era das publicações virtuais. **Comunicação e Informação**, Goiânia, v. 2, n. 1, p. 83-97, jan./jun. 1999.

ROBREDO, J. A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. D. (Org.). **Estudos Avançados em Ciência da Informação.** Brasília, DF: Associação dos Bibliotecários do Distrito Federal, v. 1, p. 235-274, 1982.

SILVA, Antônio Carlos da. **As teorias do signo e as significações lingüísticas**. [2004]. [Texto online]. Disponível em: <http://www.partes.com.br/ed39/teoriasignosreflexaoed39.htm>. Acesso em: 10 jul. 2007.

SILVA, M. R. da; FUJITA, M. S. L. A prática de indexação: análise da evolução e tendências teóricas e metodológica. **TransInformação**, Campinas, v. 16, n. 2, p. 133–161, 2004. Disponível em: <http://revistas.puc-campinas.edu.br/transinfo/viewissue.php? id=7#Artigos>. Acesso em: 22 fev. 2008.

TAXONOMIA. In: GLOSSÁRIO NETIC. [S.l.]: Portal NETIC - Núcleo de Estudos em Tecnologias para Informação e Conhecimento, [200-]. Disponível em: <http://www.netic.com.br/ glossario.html#T>. Acesso em: 24 jul. 2007.