ARTIGO

ANÁLISE COMPARATIVA E DE CONSISTÊNCIA ENTRE REPRESENTAÇÕES AUTOMÁTICA E MANUAL DE INFORMAÇÕES DOCUMENTÁRIAS*

COMPARATIVE AND CONSISTENCY ANALYSIS BETWEEN AUTOMATIC AND MANUAL REPRESENTATIONS OF DOCUMENTARY INFORMATION

Gabriel Santos ALCAIDE**
Roberto Júlio GAVA***
Willame Santos RODRIGUES****
Débora Ferreira SANTOS*****

RESUMO

Analisa a consistência dos produtos de indexação automática em domínios científicos de Saneamento Básico e de Educação. Procura averiguar se essa via se assemelha, em resultados, aos processos de indexação que utilizam a metodologia da Análise Documentária. Busca referencial teórico na Análise Documentária, a fim de reconhecer os parâmetros norteadores da análise e síntese de textos escritos, e na Terminologia, enquanto Ciência e objeto, para compreender a organização e as características dos vocabulários dos domínios do saber. Utiliza uma representação automática baseada em um modelo estatístico-morfológico, onde a extração dos léxicos é feita com o auxílio de um dicionário que possui somente uma lista de palavras e as suas respectivas classes gramaticais. Observa que se o método vir a reconhecer significados e termos compostos, e a partir destes operar com relações/redes semânticas, o processo/produto de representação automática apresentará adequados níveis de desempenho em uma efetiva indexação automática.

Palavras-chave: *Indexação automática; Terminologia; Consistência; Indexação; Representação documentária; Método estatístico-morfológico.*

(*) Síntese do Trabalho de Conclusão de Curso – TCC, aprovado na Faculdade de Biblioteconomia e Ciência da Informação – FaBCI da Fundação Escola de Sociologia e Política de São Paulo – FESP, em dezembro de 2000, sob a orientação da Prof.a Silvia Gagliardi Rocha e a coordenação do Prof. Claudio Marcondes Filho.

(**) Bacharel em Biblioteconomia, Bibliotecário Analista da Faculdade Santa Marcelina - E-mail: gsalcaide@uol.com.br
(***) Bacharel em Biblioteconomia, Chefe da Seção de Manuscritos Arquivo da Cúria Metropolitana de São Paulo - E-mail: robertogava@yahoo.com
(****) Bacharel em Biblioteconomia, Encarregado de Biblioteca Sindicato dos Contabilistas de São Paulo, Auxiliar de Ensino na Faculdade de Biblioteconomia e Ciência da Informação - 11-4186-6482 res./11-3224-5115 com./11-223-2390 fax - E-mail: willame@brfree.com.br
(*****) Bacharel em Biblioteconomia, Bibliotecária da Universidade Nove de Julho - UNINOVE - E-mail: debbsan@yahoo.com

ABSTRACT

It analyzes the consitency of the products of automatic indexation in scientific domains of basic sanitation and education. It intends to verify whether this procedure is similar, in its results, to the indexation processes that use documentary analysis methodology. It seeks theoretical reference in documentary analysis, with the purpose of recognizing the guiding parameters of analysis synthesis of written texts and in the terminology as a science and object to understanding the organization and the characteristics of the vocabularies of the knowledge domains. It uses an automatic representation, based in an statistic-morphological model, where the extraction of the words is made with the help of a dictionary that is constituted solely by a list of words and their respective grammar classes. It notes that, if the method recognizes meanings and compound terms in semantic relations and nets, the procedure/product of automatic representation will reach satisfactory levels of achievement in an effective automatic indexation.

Key-words: *Automatic indexation; Terminology; Consistency; Indexation; Documentary representation; Statistic-morphological method.*

INTRODUÇÃO

A representação via linguagens documentárias - LD's é uma das condições para se transferir informações e, portanto, mediar a comunicação entre produtores e consumidores de informações e as bases de dados. O uso do computador como um instrumento aplicado à Análise Documentária é tema de discussão deste artigo. Pretendeu-se verificar qual era o desempenho dos sistemas automáticos de indexação, em texto integral, em dois domínios científicos: o de Saneamento Básico e o de Educação. O primeiro pertence à área de Exatas e o segundo, à área de Ciências Humanas. O tema justifica-se como tentativa de averiguar como são tratados os documentos eletrônicos em ambientes da Internet e quais os resultados da indexação automática em domínios onde há ou não fixação terminológica, e se essa via se assemelha, em resultados, aos processos de indexação que utilizam a metodologia da Análise Documentária, e em qual área este tipo de representação pode vir a ser aplicado.

Mas, como preservar a contextualização de um termo que faz parte de um discurso[1] que foi indexado automaticamente? No âmbito terminológico de uma área técnica essa descontextualização estaria diminuída? Pois sabemos que o processo de extração não resolve o problema da contextualização. Não basta usar termos da terminologia como condição de referência a objetos. É preciso articulá-los em rede (relações) para que constituam um sistema de significação (TÁLAMO, 1997, p. 4). A resposta poderia ser a extração de sintagmas? Se a indexação automática enquanto produto documentário resultante da análise de textos/discursos científicos atribui léxicos de uma terminologia, como denominar este estudo: Processamento de Linguagem Terminológica ou Natural?[2] Na tentativa de obter respostas possíveis reportaremo-nos à Análise Documentária, por ser a disciplina que norteia os procedimentos de análise e a síntese de textos/documentos com vistas à representação via linguagens formalizadas.

A Análise Documentária (AD), enquanto disciplina integrante do domínio da Ciência da Informação (CI), é definida por GARDIN como "um conjunto de procedimentos seguidos para expressar o conteúdo de documentos científicos sob formas destinadas a facilitar sua recuperação ou consulta" (GARDIN, 1981, p. 29). Por permitir transcender a noção de documento textual e representar informações independente do suporte, nosso objeto de trabalho são

(1) Segundo CINTRA, existe uma variação significativa entre texto e discurso e diz respeito à força de inter-relação entre componentes lingüísticos e extralingüísticos. A definição dos lingüísticos é tida como o "conjunto de referentes textuais, composto por palavras, frases, períodos, parágrafos, capítulos, partes do texto" e os componentes extralingüísticos "os referentes situacionais que envolvem o próprio contexto situacional onde se dá o texto/discurso" (CINTRA, 1994, p. 1).

(2) A partir da observação de Gardin, inferimos esta denominação de processamento de linguagem terminológica, porque ele diz que "Análise documentária (AD) trabalha com textos científicos, e por esta razão, questiona se a transformação realizada não seria de **LE** - linguagens de especialidades da ciência para **LD** - linguagem documentária (GARDIN, 1981, p. 32-36). De acordo com Gardin, existe uma distinção entre os textos científicos produzidos no domínio das Ciências Exatas e Naturais dos que são produzidos no domínio das Ciências Humanas: "as ciências exatas e naturais possuem constituintes lexicais e estruturas textuais mais distintas da linguagem natural do que as ciências humanas (Ibid, p. 32-36).

textos escritos em formato eletrônico acessível por computador, ou, adotando a terminologia contemporânea, documentos eletrônicos.

Os métodos para a AD são tema de investigação de vários autores na tentativa de impor rigor às atividades com fins documentários. Os métodos são expostos através da identificação de operações/etapas. As fases indicadas por CUNHA são a de análise e a de síntese, onde a análise "visa, primeiramente, identificar a organização metodológica do discurso do autor/produtor através da segmentação do texto" (CUNHA, 1990, p. 73). A síntese tem como premissa "chegar a conceitos/ palavras-chave capazes de traduzir o conteúdo do discurso analisado. Procede-se então, primeiramente, a uma seleção e, depois, a uma fixação desses conceitos/ palavras-chave" (ibid., 1990, p. 76). Segundo AMARO, Cunha afirma que na análise se faz a leitura do texto, ou a contextualização e a segmentação, ou se identifica a informação principal do texto; e na síntese, realiza-se a passagem da linguagem natural (LN) para uma linguagem documentária (LD) (AMARO, 1991, p. 6).

As operações propostas por KOBASHI mostram que primeiro os textos são desestruturados, e posteriormente, as informações selecionadas são estruturadas, ou seja, parte-se à elaboração de informações documentárias. A figura abaixo mostra as operações segundo KOBASHI:

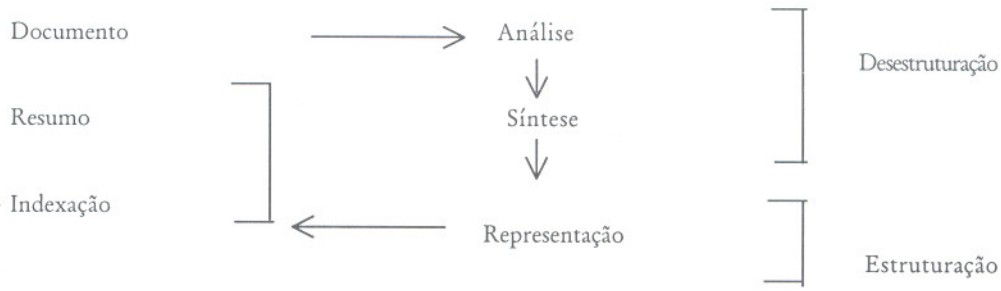

Figura 1: Fabricação da Informação Documentária (KOBASHI, 1995, p.11)

A representação para LARA é o "resultado das operações de análise e síntese do conteúdo cujo objetivo é a transferência da informação" (ibid, 1999, p. 136). KOBASHI, baseada na definição de Borko e Bernie, diz que "a indexação é o processo de analisar o conteúdo informacional dos registros do conhecimento e sua expressão na linguagem do sistema de indexação" (KOBASHI, 1995, p. 29).

O objetivo da AD é a recuperação da informação e a inserção do usuário no processo de AD, possibilita construir um conceito que realmente pode abranger todas as etapas envolvidas e o fim a que se destina a análise documentária: "a informação (...) só tem existência a partir de sua apropriação pelo usuário quando este identifica 'do que se fala' e 'de onde se fala' (lugar de sentido), a partir de suas experiências e necessidades de conhecimento (LARA, 1999, p. 8).

Um dos requisitos para se recuperar informações documentárias é o acréscimo do elemento interpretante. LARA reconhece este elemento e enfatiza que "enquanto signos documentários, as representações documentárias não podem ser vistas numa relação simples entre objeto e representação, mas supondo uma relação triádica que compreende objeto, representação e interpretante" (LARA, 1999, p. 9). Tendo como base o estabelecimento da comunicação nesse processo, a atribuição de uma linguagem de termos livres ou controlados é o que media essa comunicação usuário-sistema. Portanto, a representação, como atividade, desenvolve-se no universo da linguagem, o que permite uma comunicação[3] (ibid, 1999, p. 161).

As informações são construídas para efetuar a comunicação do conhecimento, a transferência. As LDs funcionam como instrumentos dessa ação comunicativa, na qual estão envolvidas a mediação e tradução deste conhecimento/informação para o interpretante (NOVELLINO, 1996, p. 43). Entendemos, assim como LARA, que o estudo da representação se desenvolve no universo da linguagem (LARA, 1999, p. 54) e a transferência de informação pode ser

[3] A comunicação neste âmbito é denominada comunicação documentária: "processo que envolve a codificação e a decodificação de conteúdos informacionais, ou seja, o tratamento e a recuperação da informação" (LARA, 1993, p. 72, grifo nosso). A comunicação se estabelece ao se aproximar o objeto que se quer representar do sujeito que irá interpretar esta relação (ibid, 1993, p.73).

caracterizada como uma ação comunicativa. A função principal da linguagem é instaurar a comunicação e, a seu modo, o processo de representação documentária é lingüístico-comunicacional. A linguagem documentária é que funciona como mediador entre produtores e consumidores de informação e as bases de dados. GOMES também afirma que o processo de comunicação não se estabelece "apenas entre um pólo emissor e outro receptor, mas também por canais de transferência que interagem como mediadores da ação comunicativa (GOMES, 2000, p.64). De acordo com a autora, "a interpretação é uma ação de composição" (ibid, 2000, p. 64).

A Terminologia como Parâmetro para a Construção de Linguagens de Representação Documentárias

As linguagens documentárias (LDs), por serem construídas, são linguagens formalizadas. A fim de servirem como instrumentos de medição, transformam texto em Linguagem natural (LN) ou Linguagem especializada (LE), em produtos documentários normalizados. Mas para garantir efetividade na comunicação documentária e a recuperação do contexto, o uso da Terminologia possibilita à representação remeter a determinados sistemas de significação.

A terminologia, como meio de comunicação formal entre especialistas de um domínio, é constituída por termos monorreferenciais: cada termo tem um conceito específico e se relaciona com os demais por intermédio de sua definição, permitindo a precisão e normalização desta linguagem. Isto confere, à comunicação especializada, univocidade (LIMA, 1999, p. 31). Em virtude da terminologia trabalhar com palavras em funcionamento, é possível delimitar seus valores e sua significação dentro do universo onde ocorrem, ou seja, possibilitar à representação remeter a sistemas de significação nos textos (LARA, 1993, p. 76).

A Terminologia é "um campo interdisciplinar e transdisciplinar e envolve não só a descrição e o ordenamento do conceito (nível cognitivo), mas também a transferência de conhecimento (nível comunicacional). Seus elementos centrais são conceitos e termos" (ISO 704). As terminologias enquanto objetos concretos são um "conjunto de termos que representam um sistema de conceitos de um domínio particular" (ISO 1087).

Sager, citado por DIAS, afirma que o termo terminologia é polissêmico, apontando para três concepções diferentes acerca dos termos. Como teoria, a Terminologia "é um conjunto de premissas, argumentos e conclusões necessário para explicar o relacionamento entre conceitos e termos especializados" (SAGER apud DIAS, 2000, p. 90). Como prática se define como um conjunto de métodos e atividades que visam a coleta, descrição, processamento e apresentação de termos. Enquanto produto, constitui-se por "um conjunto de termos, ou vocabulário, de um determinada especialidade" (ibid, 2000, p. 90)

Termos, Conceitos e Vocabulários dos Domínios do Saber

De acordo com a norma ISO 1087, o termo é a "designação de um conceito por meio de uma unidade lingüística definida numa língua de especialidade" (ISO 1087). Enquanto produto de uma relação extralingüística que parte do objeto, é mediado pelo conceito (LARA, 1999, p. 124). O objeto é o elemento passível de ser percebido e pode ser representado por um termo ou símbolo (ISO 1087).

Para HERMANS, o "termo científico é uma unidade lexical definida num texto científico e a condição para ser termo é a de ser definido no discurso científico". Disso resulta que "todo termo deve ser monossêmico, unívoco, prescritivo" (HERMANS, 1989, p. 529). Hermans afirma a existência de dois tipos de termos nos vocabulários científicos: os termos técnicos e os termos teóricos.

> *"Os termos técnicos designam observações, medidas, experiências, instrumentos (...) Para constituir a terminologia de um domínio técnico parte-se dos objetos preexistentes aos termos. O termo não é definido a partir de seu funcionamento no discurso, mas como correspondendo a uma noção particular dentro do campo nocional" (ibid, p. 529)*

Já os termos teóricos não se relacionam às noções preexistentes ou a representações mentais a objetos concretos e abstratos. Estes termos são utilizados com uma ou várias significações. A significação destes termos depende de seu funcionamento no contexto e das relações que ele mantém com os outros termos no enunciado (id., 1989, p. 529-530).

A significação gerada continuamente dos termos científicos é dada pelo uso que os cientistas fazem dos termos e pela forma como tal uso é assimilado pelos outros. Isso pode ocasionar termos imprecisos. No entanto, a estabilidade nas ciências é comparada à estagnação. Os cientistas procuram uma instabilidade próxima entre a indeterminação e a determinação, pois se os conceitos científicos forem muito determinados,

eles não podem funcionar como instrumentos de descoberta ou explicação (id., 1989, p. 530).

Segundo LARA, "a referência dos termos na Terminologia é formulada através de uma operação de definição" (LARA, 1999, p. 120). A definição, de acordo com a norma ISO 1087, é um "enunciado que descreve um conceito e que permite diferenciá-lo de outros conceitos no interior de um sistema de conceitos" (ISO 1087). Definir os termos teóricos implica em especificar as diferentes significações que podem ter estes termos: "especificações cronológicas, por escola de pensamento ...". E propor uma definição teórica "equivale a propor a aceitação de uma teoria e (...) as teorias são notoriamente discutíveis". Com isso é possível afirmar, segundo COPI, "que uma definição é substituída por outra à medida que nosso conhecimento e compreensão teóricos aumentam" (COPI, 1978, p. 117).

INDEXAÇÃO AUTOMÁTICA

No final da década de 50, Luhn desenvolve o índice Keyword in Context (KWIC), baseado em linguagem natural[4], que retirava termos dos títulos dos trabalhos comparando com um índice de palavras proibidas. Caso não houvesse coincidência, o termo seria admitido como palavra-chave. Essa "novidade", para GOMES, era resultado não "do aproveitamento das palavras-chave presente nos títulos dos documentos, mas da rapidez com que a tarefa poderia ser executada" (GOMES, 1989, p. 166). Mais tarde, percebeu-se que outras palavras dos trabalhos poderiam constituir palavras-chave, acrescentando-se desta forma os subtermos.

De acordo com GARCIA GUTIÉRREZ, nos anos 60 surgiram "modelos de representação automática" baseados na leitura seqüencial do texto; o autor afirma que em tais modelos predominam a representação sobre a análise (GARCIA GUTIÉRREZ, 1992, p. 33).

Segundo MAMFRIM, a mecanização do processo manual de indexação, no todo ou em parte, visa "a estabelecer rotinas que reduzam a interferência da subjetividade do indexador, tanto na análise do documento, quanto na seleção dos termos significativos" (MAMFRIM, 1991, p. 191).

ROBREDO afirma que a comparação de "cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, conduz por eliminação, e considera as palavras restantes dos textos como palavras significativas" (ROBREDO, 1982, p. 236).

O processo de indexação automática que opera com base na constituição de dicionários e anti-dicionários de "palavras vazias invariáveis" e "raízes de palavras não significativas"[5], segundo KOBASHI, "inspira-se no distribucionalismo". Em virtude de se basear em "critérios semânticos estatísticos", KOBASHI considera que esta é uma metodologia de indexação viável apenas em áreas cuja terminologia esteja estabilizada (KOBASHI, 1995, p. 30).

Com base nas afirmações de Chaumier, para CASTILHO, **a indexação automática opera com texto integral e realiza de forma automática todas as etapas da indexação, inclusive as etapas de análise e síntese** (CASTILHO, 1995, p. 12, grifo nosso).

ROLE, opondo-se à indexação automática, diz que os sistemas informatizados não realizam nenhuma espécie de análise, mas trabalham somente com a extração de palavras do texto (ROLE, 1993, p.140).

Através da revisão de literatura, pode-se verificar que a mecanização do processo manual de indexação somente é realizado em parte, em específico na fase de síntese, ou seja, na atribuição de léxicos documentários, portanto, com a prepoderância da representação sobre a análise.

Afirmar hoje que a indexação é um processo subjetivo é inadequado, por haver um campo que propõe processos metodológicos de análise e síntese: a AD. Poderá haver subjetividade quando não for empregado um parâmetro metodológico. Entretanto, a extração seria capaz de evitar a subjetividade? Os teóricos abordam a existência de subjetividade, mas não esclarecem/propõem uma sistematização analítica.

Métodos Estatístico-Morfológicos Aplicados a Textos Escritos

Os modelos estatístico-morfológicos partem da premissa de que "as palavras de um texto dividem-se em duas categorias: aquelas que são portadoras de uma significação e as demais" (COULON & KAYSER,

(4) A linguagem natural é "sinônimo de discurso comum, isto é, a linguagem utilizada habitualmente na escrita e na fala, e que é o contrário de vocabulário controlado" (LANCASTER, 1993, p. 200).

(5) O dicionário de "palavras vazias invariáveis" é constituído de preposições, conjunções, advérbios, entre outros. O outro, de "raízes de palavras não significativas" na área de conhecimento processada (KOBASHI, 1995, p. 30).

1992, p. 18). Os modelos dividem-se em análise estatística e análise léxico-morfológica.

A análise estatística é uma indexação baseada em cálculo freqüencial automático: "consiste em extrair os termos de um texto e contabilizar suas repetições" (GARCIA GUTIÉRREZ, 1992, p. 133). Desta forma, pode-se comparar sua freqüência com a de outros termos do mesmo texto ou de outros.

Luhn, ao criar o Keyword in Context (KWIC), utilizou o cálculo freqüencial automático centralizando-se no fato de que "a freqüência de determinadas palavras em um texto dá a medida da representatividade destas palavras no mesmo" (COYAUD, 1967, p. 11).

COYAUD afirma que este método "não considera possíveis sinonímias ou polissemias presentes no texto" (Ibid, p.14). À partir desta constatação, GARCIA GUTIÉRREZ propõe a utilização deste método somente a textos com "estabilidade terminológica", observando que é "tarefa impossível em grande quantidade de textos, especialmente os que contêm informações sobre atualidades ou Ciências Sociais" (GARCIA GUTIÉRREZ, 1992, p.133).

A análise léxico-morfológica iniciou-se também com os estudos de Luhn, e visa solucionar as falhas que ocorrem devido à recuperação de palavras não significativas"[6] através da utilização de antidicionários, compostos por palavras sem significância, também chamadas de palavras vazias; comparam-se as palavras extraídas do texto com este antidicionário e, caso haja coincidência, são desconsideradas na elaboração do índice (LANCASTER, 1993, p.48).

ROLE destaca, em seu trabalho sobre sistemas, o critério de substituir os termos do texto por outros julgados mais adequados, "ou seja, as palavras extraídas do texto são comparadas com um vocabulário controlado (...) possibilitando o controle de sinonímias e polissemias" (ROLE, 1993, p.138).

A partir do momento em que se identificam, nos textos, os termos, pela análise estatística, o "analisador morfológico consiste em achar (...) a forma representativa" destes termos "armazenada no léxico". Esta forma representativa, "como nos dicionários, conserva-se em forma única (...) devendo encontrar-se as demais por meio de regras que descrevem as flexões possíveis" (COULON & KAYSER, 1992, p.40).

METODOLOGIA

Foram selecionados quatro textos dentro dos domínios científicos Educação e Saneamento básico, que correspondem respectivamente às áreas de Ciências Humanas e de Ciências Exatas. Sendo 3 científicos e 1 de divulgação para cada domínio. O conjunto se apresenta:

FARIA FILHO, Luciano Mendes de. O espaço escolar como objeto da história da educação : algumas reflexões. **Revista da Faculdade de Educação** [online], São Paulo, v.24, n.1, jan./jun. 1998. Disponível na internet: <http://www.scielo.br/cgi-bin/fbpe/fbtext?got=last&pid=S0102-25551998000100010&lng=pt&nrm=isso>

KAWASAKI, Clarice Sumi. Universidades públicas e sociedade : uma parceria necessária. **Revista da Faculdade de Educação [online]**, São Paulo, v.23, n.1-2, jan./dez. 1997. Disponível na internet: <http://www.scielo.br/cgi-bin/fbpe/fbtext?got=last&pid=S0102-551997000100013&lng=pt&nrm=isso>

SILVA, Luiz Carlos Faria da. Possíveis incompletudes e equívocos dos discursos sobre a questão da disciplina. **Educação & Sociedade** [online], Campinas, v.19, n.62, abr. 1998. Disponível na internet: <http://www.scielo.br/cgi-bin/fbpe/fbtext?got=last&pid=S0101-301998000100007&lng=pt&nrm=isso>

ZENTI, Luciana. A arte de ser professor. **Nova Escola online** [online], out. 2000. Disponível na internet: <http://www.uol.com.br/novaescola/>

MOTA, Suetônio, BEZERRA, Francisco Cesar, TOMÉ, Luciana Mota. A avaliação do desempenho de culturas irrigadas com esgoto tratado. In: CONGRESSO BRASILEIRO DE ENGENHARIA SANITÁRIA E AMBIENTAL, 20, Rio de Janeiro, 1999. **Anais ...** [online]. Rio de Janeiro : ABES. 1999. Disponível na internet: <http://www.saneamentobasico.com.br/materia/estudos/files/textos97/i-003.doc>

LIMA, Márcio Rogério Azevedo, REALI, Marco Antonio Penalva. Tratamento físico-químico das águas residuárias de uma indústria de papel utilizando-se a flotação por ar dissolvido. In: CONGRESSO BRASILEIRO DE ENGENHARIA SANITÁRIA E AMBIENTAL, 20, Rio de Janeiro, 1999. **Anais ...** [online]. Rio de Janeiro : ABES. 1999. Disponível na internet: <http://www.saneamentobasico.com.br/materia/estudos/files/textos97/i-006.doc>

COSTA, Alberto José Moitta Pinto da, *et all.* Estudo de tratabilidade de água residuária sintética simulando despejo líquido de coquerias. In: CONGRESSO BRASILEIRO DE ENGENHARIA SANITÁRIA E AMBIENTAL, 20, Rio de

[6] Palavras não significativas, conforme Lancaster, são "artigos, preposições, conjunções e assemelhados" (ibid, 1993, p.48).

Janeiro, 1999. **Anais ...** [online]. Rio de Janeiro : ABES. 1999. Disponível **na** internet: <http://www.saneamentobasico.com.br/materia/estudos/files/textos97/i-060.doc>

LUCÍRIO, Ivonete D. Parados e sufocados. **Superinteressante online** [online]. jun. 1996. Disponível na internet: <http://www2.uol.com.br/super/super0696/polu.html>

Para o processo de indexação manual e automática[7] foram utilizadas as seguintes linguagens documentárias:

VIEIRA, Maria da Graça Camargo (org.). **Vocabulário controlado [de educação]**. São Paulo: Fundação Carlos Chagas, Biblioteca Ana Maria Poppovic, Departamento de Pesquisas Educacionais, 1998.

VOCABULÁRIO Controlado [de saneamento básico]. 2 ed. São Paulo : Sabesp, TDST, 1997.

Para a indexação automatizada foi utilizado o protótipo[8] de FERNEDA.

Os termos compostos dos vocabulários controlados foram fatorados, quanto à forma sintática, em termos simples. Estes, com os já existentes no vocabulário, foram alfabetados e suas duplicatas excluídas, mantendo-se somente uma unidade de cada palavra.

Com as palavras organizadas em ordem alfabética e de posse de dicionários de língua portuguesa foram atribuídas as categorias gramaticais às palavras da relação. Aos nomes próprios e termos em língua estrangeira foi atribuída a classe substantivo. Às palavras pertencentes a mais de uma classe gramatical foi atribuída a classe na qual a palavra apresenta maior significação no universo estudado.

Os termos, que são a forma canônica[9] das palavras, foram normalizados para masculino-singular no caso de substantivos e adjetivos e na forma infinitiva para os verbos. Com as classes gramaticais atribuídas às palavras procedeu-se à alimentação do dicionário que o protótipo utiliza para o processamento dos textos.

Todos os artigos foram coletados via internet, sem exceção, através de ferramentas de busca da rede e pelos serviços disponíveis aos assinantes de provedores de acesso.

Os artigos foram gravados em um computador local (*off line*) na forma integral de seu conteúdo apresentado na rede. Trabalhando em modo local (*off line*), os artigos foram abertos em um processador de textos, o Winword.

O título do periódico bem como as demais informações deste, o título do artigo, o resumo, as palavras-chave, biografias dos autores, ilustrações, tabelas, gráficos, legendas e referências bibliográficas foram apagados deixando-se somente os títulos dos capítulos e o texto na íntegra. Através do mesmo processador de textos foi realizada a contagem das palavras dos artigos. Estes foram gravados com a extensão que o protótipo é capaz de reconhecer[10].

O protótipo trabalha em três fases distintas, mas interdependentes:

- Primeiramente são extraídos do texto os termos significativos existentes em seu dicionário e é feita a contagem total de termos. Aos termos é relacionado o número do texto que está sendo indexado, já que podem existir vários consecutivamente, os números do parágrafo em que aparece o termo e sua posição neste parágrafo;
- No segundo passo o protótipo cria uma matriz relacional entre os termos, onde são avaliadas a freqüência que um termo aparece no texto e sua relação com os demais termos;
- Por último o protótipo extrai as palavras-chaves baseando-se nas forças relacionais calculadas na matriz gerada no passo anterior (FERNEDA, 1997, p. 47-58).

Para a indexação manual os textos foram impressos e enumerados. Excluíram-se o resumo e as palavras-chave. Os vocabulários controlados foram entregues aos indexadores especialistas nas áreas dos domínios tratados.

RESULTADOS E ANÁLISE COMPARATIVA E DE CONSISTÊNCIA DOS PROCESSOS E PRODUTOS DOCUMENTÁRIOS

O protótipo identificou uma quantidade de termos significativos nos textos relativamente baixa: 23,50% no texto 1, 17,00% no texto 2, 19,13% no texto 3, 16,83% no texto 4, 23,41% no texto 5, 21, 47% no texto 6, 17, 76% no texto 7 e 13,27% no texto 8.

(7) O protótipo consegue efetuar a extração de palavras através da comparação a um dicionário morfológico. Para possibilitar o processo alimentamos o programa com os vocabulários citados.
(8) FERNEDA nos cedeu gentilmente este protótipo.
(9) Referimo-nos ao verbete propriamente dito.
(10) O protótipo trabalha com arquivos de extensão txt.

Estando, portanto, entre a faixa de 10% e 25%, com valor médio de 19,05%.

Os textos de divulgação, 4 e 8, foram os que apresentaram menor índice de ocorrência de termos significativos, obtendo 16,83% e 13,27% respectivamente. Estes percentuais estão claramente identificados com os textos que possuem menor quantidade de palavras, 2.697 e 1.364 tanto na área de Exatas como na de Humanas.

Os maiores índices não demonstraram o mesmo comportamento, não estando relacionados aos textos com maior quantidade de palavras. Este fato pode ser atribuído a um maior domínio terminológico do autor dos artigos, estando mais consciente de sua área de atuação, como também estar relacionado a um dicionário morfológico[11] alimentado no sistema com maior ou menor abrangência.

As palavras-chave atribuídas somente apresentaram uma relação clara quanto a suas quantidades aos textos de divulgação, textos 4 e 8, tendo ocorrido os menores índices atribuídos, 5 e 2, respectivamente. Nos demais textos as quantidades não puderam ser avaliadas por ocorrerem de forma aleatória, tanto em relação a quantidade de palavras quanto a quantidade de termos significativos identificados.

Observamos uma quantidade média de palavras alta nos textos da área de humanas, 5.397 palavras aproximadamente, em comparação a área de exatas, 1.385 palavras.

Numericamente a indexação automática mostrou-se extensiva com relação a indexação manual. Novamente a exceção é feita aos textos de divulgação, 4 e 8, que apresentam valores baixos e muito próximos para as duas formas de indexação e no caso do texto 8, na área de exatas, a quantidade de palavras-chave atribuídas pela indexação manual foi maior que a automática, como pode ser observado no quadro a seguir:

	PALAVRAS-CHAVE	PALAVRAS-CHAVE
TEXTOS	INDEXAÇÃO AUTOMÁTICA	INDEXAÇÃO MANUAL
1	47	6
2	36	5
3	31	5
4	**5**	**4**
5	14	7
6	27	4
7	10	5
8	**2**	**4**

Quadro 1: Quantidade de palavras-chave empregadas por ambas as indexações

A título de complementação foi feita uma comparação entre os termos de indexação utilizados na representação dos textos na internet e os resultados obtidos nesse trabalho, para averiguar, mesmo não tendo conhecimento acerca do uso ou não de uma LD, se os produtos se assemelham aos automáticos ou aos manuais.

A indexação "original" aproxima-se do que foi realizada pelo processo manual: uso da generalidade, significado ambíguo; mas em determinados contextos, os léxicos documentários remetiam com pertinência ao sistema de significação dos discursos.

Salientamos que para os nossos modelos de indexação não se determinou uma política de indexação e que para a indexação "original" esta política é desconhecida.

Os resultados no Domínio de Educação foram:

Texto 1: Universidades públicas e sociedade : uma parceria necessária.

✓ Total de palavras do texto : 5192 palavras

Indexação automática

✓ Foram localizados 1220 termos significativos pelo sistema.

[11] Lista de palavras com suas respectivas categorias gramaticais.

1. Ambiente conselho meio	2. Área ambiental
3. Ambiental controle	4. Ambiental estudo
5. Ambiental gestão	6. Ambiental impacto
7. Ambiental licença	8. Ambiental questões
9. Aspecto econômico	10. Econômico social
11. Econômico conhecimento valor	12. Econômico globalização
13. Atividade pesquisa	14. Atividade obra
15. Pesquisa científica educação	16. Pesquisa centro
17. Pesquisa básica educação	18. Pesquisa universidade
19. Pesquisa aplicado[12]	20. Captação científico tecnológico
21. Científico social	22. Científico conhecimento
23. Científico patrimônio	24. Social demanda
25. Social problema	26. Conhecimento produção
27. Desenvolvimento política	28. Desenvolvimento modelo
29. Desenvolvimento projeto	30. Política pública
31. Educação superior	32. Espaço público
33. Formação profissional	34. Gestão participativa
35. Globalização processo	36. Igualdade condição
37. Instituição educativo	38. Lei mercado
39. Novas tecnologia	40. Papel universidade
41. Universidade pública	42. Universidade projeto
43. Universidade estadual	44. Universidade sociedade
45. Projeto nacional	46. Qualidade total
47. Sistema educacional	

Quadro 2: Palavras-chave extraídas do texto 1

Indexação manual

1. Universidades públicas	4. Universidades públicas
2. Globalização	5. Globalização
3. Desigualdades sociais	6. Desigualdades sociais

Quadro 3: Palavras-chave atribuídas ao texto 1

[12] Como já dissemos, p. 25, os léxicos do dicionário morfológico que alimentamos no protótipo estão no masculino-singular.

O processo documentário manual realizado no texto 1 de educação atribuiu o termo /Meio ambiente/[13]. No entanto, ao pós-coordenar os termos /Meio ambiente e Universidades públicas/ o significado será ambíguo, por remeter a outro sentido: poderia indicar que o texto aborda sobre como é a área ambiental das universidades públicas[14]. O aspecto abordado é o do dever das universidades públicas em elaborar projetos para o meio ambiente a fim de atender às necessidades sociais. Para uma recuperação pertinente do contexto, deveria-se qualificar o termo /Meio ambiente/ com / Projetos/. No processo automático seria possível recuperar este contexto, pois o protótipo pré-coordenou[15] os termos /Universidade projeto/ e realizando a lógica booleana entre os pré-coordenados e /Meio ambiente/, recuperaria-se o sentido.

Ambos os processos reconheceram como termo /Universidades públicas/, mas o automático não o flexionou. Isso não impediria de recuperá-lo. A discussão do autor sobre as universidades públicas é abordada no contexto da globalização: este termo também foi tido como significativo, entretanto, o automático gerou os léxicos /Processo/ e /Economia/ que poderiam ser considerados como qualificadores.

Os termos /Universidades públicas/ pós-coordenado com /Sociedade/ não conseguiria resgatar o significado: poderia ser a comunidade científica ou não. Por se tratar da discussão do papel das universidades públicas na sociedade brasileira, ou se acrescentaria o termo /Papel das universidades públicas/, ou se atribuiria o termo encontrado no vocabulário /Relações universidade-sociedade/. Na automática este aspecto foi extraído sob o termo /Papel universidade/. Se fizéssemos uma pós-coordenação com este e o termo / Formação profissional/ que é relacionado ao anterior, o sentido seria recuperado. O termo relacionado foi reconhecido por nós, em virtude da discussão ser no contexto globalização e o papel da universidades públicas em formar profissionais para este mundo globalizado. O léxico /Sociedade/ sem o qualificador / Brasil/ acarretaria na perda do significado. Suponhamos que nesta base de dados alguém queira consultar sobre a sociedade e o meio ambiente, ao efetuar a lógica booleana este texto seria recuperado, mas não o contexto, não a significação construída através do discurso do autor. Percebe-se que as falsas associações ocorrem não somente na automática, por ser exaustiva, por ter uma alta revocação. As falsas associações também acontecem porque não se qualificam os termos. O mesmo ruído poderia ocorrer na comunicação documentária do termo /Economia/. Se um interpretante necessitasse acessar informações sobre a situação econômica das universidades públicas, e pós coordenasse os termos, recuperaria-se este texto[16], logo, a indexação proporcionaria uma falsa coordenação, e perda de significado.

No processo automático houve uma aproximação a um termo relacionado /Igualdade condição/ ao atribuído pelo manual /Desigualdades sociais/. Independente do automático não o ter extraído como / Desigualdades sociais/ através da relação assimilada por nós, seria perfeitamente adequado o uso do termo / Igualdade condição/.

A indexação manual gerou 6 termos, e poderíamos inferir que foi utilizado como parâmetro a precisão, no entanto, a quantidade relativamente pequena não quer dizer uso de precisão. Deve-se averiguar se os termos realmente representam os contextos dos discursos científicos. A precisão implica em fixar sentidos, como tentativa de diminuir a plurissignificação e permitir uma alta especificidade. Como os termos eram genéricos, não possuíam qualificadores, a revocação tornou-se alta e perdeu-se o significado, em virtude dos léxicos documentários não serem unívocos: recuperariam-se informações relevantes, mas não precisas.

A indexação automática foi exaustiva[17], por gerar 47 termos, teoricamente sendo exaustiva, a revocação[18] seria alta. A maior parte dos termos extraídos são sinônimos, portanto, a exaustividade ocorreu em virtude da sinonímia.

Texto 2: **Possíveis incompletudes e equívocos dos discursos sobre a questão da disciplina**

✓ Total de palavras do texto: 8.433 palavras.

(13) Adotou-se escrever o termo entre barras, porque "este artifício gráfico indica um signo assumido na sua forma significante" (ECO,1997, p. 24).

(14) Para Minsky, as ambigüidades na linguagem natural "não advém apenas do fato de que as palavras podem ser reagrupadas de diversas maneiras, mas ainda do fato de que cada palavra pode ter diferentes sentidos..." (MINSKY, 1969 apud KURAMOTO, 1995, p. 5).

(15) O protótipo aproxima-se de uma pré-coordenação.

(16) A relevância é entendida como a "capacidade do sistema em fornecer respostas (referências) que realmente correspondam à questão proposta" (CARNEIRO, 1985, p. 234).

(17) Ao possibilitar um número grande de pontos de acesso, a exaustividade tenta recuperar o âmbito de abrangência do documento. Segundo LANCASTER, "...a indexação exaustiva redundará em menor precisão das buscas...será recuperado um número maior de itens que o usuário considera como não sendo pertinentes a sua necessidade de informação" (LANCASTER, 1993, p. 23).

(18) Para CARNEIRO, a revocação "é a capacidade do sistema em assegurar a recuperação ... de documentos relevantes e a precisão se relaciona à capacidade do sistema em impedir a recuperação de documentos não-relevantes" (CARNEIRO, 1985, p. 234).

Indexação automática

✓ ✓ Foram localizados 1.441 termos significati-vos pelo sistema

1. Ação educacional	2. Ação educativo
3. Ação humano	4. Ações humano
5. Aplicado educação	6. Aplicado psicodrama
7. Educação conteúdo político	8. Educação disciplina escolar
9. Educação disciplina poder	10. Educação disciplina relação
11. Ato educativo	12. Concepção histórica
13. História homem	14. Dimensão psicológico
15. Disciplina abordagem	16. Disciplina problema
17. Disciplina questão	18. Escolar instituição
19. Poder disciplinar	20. Poder autoridade
21. Disciplinar escola	22. Escola nova
23. Escola relação	24. Escola violência
25. Nova qualidade	26. Natureza conceitual
27. Prática educacional	28. Prática educativo
29. Processo educativo	30. Professor aluno relação
31. Projeto pedagógico	32. Psicodrama teoria
33. Relação punição vigilância	34. Teoria educacional
35. Termo discurso	36. Trabalho grupo

Quadro 4: Palavras-chave extraídas do texto 2

Indexação manual

1. Normas disciplinares	4. Educação
2. Autoridade	5. Prática de ensino
3. Relações sociais	

Quadro 5: Palavras-chave atribuídas ao texto 2

Neste texto são abordados dois aspectos acerca da questão da disciplina, sendo que um classificado em formulações teóricas (psicologia) e o outro em análise crítica da realidade do poder.

Na indexação manual a atribuição do termo /Educação/ pós-coordenado com /Normas disciplinares/ perde significado, pois tenciona ao sentido de leis e regras disciplinares na educação. O autor enfoca as relações entre disciplina escolar, educação e transformação social. No processo automático todos estes sentidos foram representados, pois o protótipo pré-coordenou /Educação disciplina relação/. Ao passo que transformação social, tanto a manual quanto a automática não o consideraram como termo representativo.

A indexação automática, sendo exaustiva, extraiu termos sinônimos como /Prática educacional/ e /Prática educativo/ e seus sub-conjuntos /Aplicado psicodrama,

/Trabalho grupo/ e /Psicodrama teoria/, e a manual atribuiu somente /Prática de Ensino/. Ambos os produtos documentários conseguiriam recuperar o contexto.

O autor analisa a questão do poder relacionado à disciplina. Neste caso a manual atribuiu o termo / Normas disciplinares/ mas, em virtude de sua falta de precisão não considerou a relação com o termo poder. A automática pré-coordenou /Educação disciplina poder/ podendo-se realizar a lógica booleana entre pré-coordenados e /Relação/, o que possibilitaria a recuperação da pertinência.

O termo atribuído pelo processo manual / Relações sociais/ pré-coordenado com /Educação/ torna-se ambíguo: poderia indicar educação como apenas a integração do indivíduo na sociedade. O texto analisa a educação como transformação da sociedade. A automática também não conseguiu reconhecer estes termos, no máximo pré-coordenou /Processo educativo/ e mesmo realizando a lógica booleana entre /Relação/ e /Nova qualidade/ ainda não remeteria ao significado proposto pelo discurso.

Texto 3: **O espaço escolar como objeto da história da educação: algumas reflexões**

✓ Total de palavras no texto: 5.267 palavras.

Indexação automática

✓ Foram localizados 1.008 termos significativos pelo sistema.

1. Água potável	2. Arquitetura escolar
3. Escolar construção	4. Escolar cultura nova
5. Escolar educação	6. Escolar educativo
7. Escolar escola	8. Escolar espaço grupo
9. Escolar museu	10. Escolar relação
11. Construção novo	12. Controle professor
13. Corpo docente	14. Cultura urbano
15. Cultura primeira	16. Urbano espaço
17. Urbano mundo	18. Dimensão espacial
19. Escola instituição	20. Escola pública
21. Pública instrução	22. Espaço interno
23. Exercício físico	24. Físico mundo
25. Nova capital	26. Nova pedagogia
27. Número aluno	28. Planta tipo
29. Político econômico	30. Sala aula
31. Trabalho livre	

Quadro 6: Palavras-chave extraídas do texto 3

Indexação manual

1. História da educação	4. Arquitetura
2. Escolas	5. Espaço físico
3. Minas Gerais	

Quadro 7: Palavras-chave atribuídas ao texto 3

No processo automático para o texto 3 de educação foi extraído /Água potável/, termo este completamente fora do contexto abordado pelo autor. Pela alimentação no protótipo de duas áreas de especialidade distintas em um mesmo arquivo de termos, a possibilidade dessa ocorrência é real pelo fato do programa não criterizar conceitos, mas sim ocorrências.

É referenciado /Nova capital/, termo que realmente apresenta relevância no texto, porém, o autor se refere a Belo Horizonte como a nova capital de Minas Gerais, já que se trata de um estudo histórico. Logo, a ocorrência de /Minas Gerais/ na indexação manual é pertinente, apesar de sua atribuição não estar contida em uma situação de qualificador mas como termo atribuído.

/Cultura urbano/ e /Urbano mundo/ demonstram relações de sinonímia possíveis em um processo de indexação automatizado sem o controle de conceitos, como ocorre neste caso.

Os termos /Físico mundo/, /Espaço interno/, / Planta tipo/, /Dimensão espacial/, /Arquitetura escolar/ e /Escolar construção/ são, também, sinonímias, apesar de estarem representando o conteúdo do artigo. Isto é conferido pela extração de /Escolas/, / Arquitetura/ e / Espaço físico/ no processo manual, que em uma situação de pós-coordenação, por exemplo, entre /Escola/ e / Arquitetura/ recuperaria o conteúdo pertinente. Sendo que /Arquitetura/ é um termo relacionado à área em questão.

A atribuição, na indexação manual, do descritor /História da educação/ apresenta-se como termo geral necessário, o que não ocorre na indexação automática.

Texto 4: A arte de ser professor

✓ Total de palavras no texto: 2.697 palavras.

Indexação automática

✓ Foram localizados 454 termos significativos pelo sistema.

1. An professor	2. Energia elétrico
3. Ensino fundamental	4. Escola pública
5. Zona rural	

Quadro 8: Palavras-chave extraídas ao texto 4

Indexação manual

1. Relação professor-aluno	3. Comunidade
2. Prática de ensino	4. Professores

Quadro 9: Palavras-chave atribuídas ao texto 4

O termo /Relação professor-aluno/ atribuído pelo processo manual é pertinente ao assunto abordado no texto. No entanto, no automático não se conseguiu gerar essa particularidade. Tanto /Prática de ensino/ conseguiria remeter ao contexto, quanto /Comunidade/ , por meio de pós-coordenação. Entretanto, o termo / Professores/ é inadequado por ser genérico, pois o texto trata sobre a satisfação profissional através do relato de experiências. Para resolver este problema, seria necessário utilizar um qualificador /Professor (relato de experiências)/ ou /Professor (satisfação profissional)/

No processo de indexação automática, o protótipo reconheceu os termos /Ensino fundamental/ e /Escola pública/, mas para recuperar a pertinência e a representação do sentido, construído pelo discurso do autor, é necessário efetuar uma pós-coordenação entre os citados, /Zona rural/ e o que indicamos no parágrafo anterior.

/Ano professor/ não tem representação lógica; foi extraído, pela freqüência que estas palavras aparecem no texto; entretanto, mesmo reconhecendo a freqüência, o método não possibilita o reconhecimento da significância. É abordado no texto a satisfação profissional , através do relato de professores em diversas regiões do Brasil.

/Energia elétrico/ foi extraído pelo fato de alguns professores terem tido experiências em lecionar em comunidades onde não havia energia elétrica. Foi

abordado superficialmente, sendo assim, não seria adequado considerá-lo como termo significante.

Os de Saneamento Básico podem ser observados abaixo:

Texto 5: **A avaliação do desempenho de culturas irrigadas com esgoto tratado**

✓ Total de palavras no texto: 1.512 palavras.

Indexação automática

✓ Foram localizados 354 termos significativos pelo sistema.

1. Abastecimento humano	2. Água esgoto
3. Água irrigação	4. Água curso
5. Água reuso	6. Esgoto característica
7. Esgoto Doméstico tratado	8. Esgoto estação tratamento
9. Irrigação sistema	10. Efluente estação tratamento
11. Engenharia sanitário	12. Sanitário ambiental
13. Matéria seca	14. Seca proteína bruta

Quadro 10: Palavras-chave extraídas do texto 5

Indexação manual

1. Esgoto doméstico	5. Reuso da água
2. Irrigação	6. Reutilização de esgoto tratado
3. Irrigação com esgoto tratado	7. Tratamento de esgoto
4. Lodo ativado	

Quadro 11: Palavras-chave atribuídas ao texto 5

No texto 5 de Saneamento básico, a indexação manual gerou como léxico /Esgoto doméstico/, mas o ponto de vista privilegiado é o do Tratamento de esgoto doméstico, e este aspecto foi reconhecido como significativo pelo processo automático sob o termo / Esgoto doméstico tratado/, mesmo que não esteja normalizado este produto, seria possível recuperar a relevância desta informação. A manual atribuiu somente /Tratamento de esgoto/, no entanto isso geraria uma ambigüidade, pois este esgoto tratado é o doméstico ou o industrial? E como já foi explicitado na proposição anterior, o protótipo conseguiu captar este aspecto. Outro termo relacionado, reconhecido por nós, foi considerado como produto /Esgoto estação tratamento/ , já a manual não conseguiu captar essa nuance, entretanto, é necessário refletir que o tratamento de esgoto doméstico ou industrial somente poderá ser realizado a partir de uma estação de tratamento, com a aplicação de técnicas e equipamentos.

Observamos que o protótipo levantou um conceito /Água reuso/, mesmo estando pré-coordenado seria possível recuperá-lo com pertinência. O meio manual gerou uma sinonímia ao atribuir /Reuso da água/ e /Reutilização de esgoto tratado/, no entanto, o automático somente reconheceu um dos termos, o qual já tínhamos dito que é um conceito. Pós-coordenando / Água reuso/ e /Irrigação sistema/ ou /Esgoto doméstico tratado/ e /Irrigação sistema/ seria alcançada a precisão e o contexto do discurso. Insistimos em enfatizar que a exaustividade dos produtos documentários automáticos ocorreu em virtude da sinonímia, mas neste domínio técnico, foi gerada também porque o protótipo reconheceu termos genéricos e sem representatividade como /Curso água/ e /Abastecimento humano/, em virtude de serem tratados de forma secundária.

A princípio poderíamos considerar os léxicos que o protótipo extraiu /Matéria seca/ e /Seca proteína bruta/ como assuntos com leve menção, no entanto,

suponhamos que um interpretante, para fundamentar seu projeto de pesquisa, deseje encontrar um estudo de irrigação com esgoto tratado que utilizou como características agronômicas de culturas a proteína bruta e a massa seca. Imaginar este alto grau de especificidade na transferência implica em considerar como relevantes determinados casos de secundarismo em léxicos documentários.

Se a indexação automática não fosse exaustiva em virtude da sinonímia iria se aproximar potencialmente da manual. Foi possível averiguar que em determinadas representações, a automática conseguiu remeter aos significados dos discursos e a manual em outras, necessitava ser precisa, não como referência a fins quantitativos, mas em aspectos semânticos. Isso ocorreu provavelmente pelo uso de uma lista alfabética de assuntos, e segundo LARA "nestas listas as palavras não são portadoras de significado porque este é remetido às várias possibilidades de sentido registradas pelo léxico: ao significar potencialmente tudo, acabam por não significar nada" (LARA, 1999, p. 62).

Texto 6: **Tratamento físico-químico das águas residuárias de uma indústria de papel utilizando-se a flotação por ar dissolvido**

✓ Total de palavras no texto: 2.401 palavras.

Indexação automática

✓ Foram localizados 522 termos significativos pelo sistema.

1. Amostra bruta	2. Amostra flotação
3. Amostra pH	4. Ar dissolvido
5. Engenharia sanitário ambiental	6. Engenharia sanitário brasileiro
7. Congresso brasileiro	8. Cloreto férrico ensaio
9. Férrico remoção ensaio	10. Dosagem coagulante
11. Dosagem ensaio	12. Dosagem valor
13. Coagulante pH	14. Eficiência flotação
15. Eficiência remoção	16. Remoção resultado ensaio
17. Resultado tabela	18. Ensaio flotação
19. Ensaio pH	20. Ensaio amido
21. Recirculação volume	22. Volume água
23. Variação	24. Medida
25. Menores	26. Valor pH
27. Tabela nota	

Quadro 12: Palavras-chave extraídas do texto 6

Indexação manual

1. Flotação por ar dissolvido	
2. Reuso da água	3. Tratamento de efluentes de industrias de papel
4. Tratamento físico-químico	

Quadro 13: Palavras-chave atribuídas ao texto 6

A indexação manual do texto 6, da área de Saneamento Básico, não conseguiu resgatar, de forma até mesmo genérica, tratamento de resíduos industriais. A indexação automática extraiu 26 palavras-chave, enquanto a manual atribuiu apenas 4.

O autor relata neste estudo sobre o tratamento de resíduos industriais a partir da técnica de flotação por ar dissolvido, em uma empresa de papel e embalagens, utilizando dosagens de alguns materiais para a recirculação da água.

Ao propor uma leitura documentária do texto, notamos a preocupação do autor com o meio ambiente. Neste sentido a manual não reconheceu nenhum termo. O automático pré-coordenou /Engenharia sanitário ambiental/ além de reconhecer que tal estudo é feito no Brasil, gerando o léxico /Brasileiro/.

Tanto na manual, a técnica utilizada pelo autor /Flotação por ar dissolvido/, quanto na automática /Flotação/ e /Ar dissolvido/ foram reconhecidos como termos significativos.

O léxico /Recirculação/ pós-coordenado com /Água/ conseguiria recuperar o sentido, no entanto, é sinônimo de /Reuso da água/, o qual foi atribuído pela manual.

A indexação automática foi exaustiva, extraiu termos que não possuem relevância por se tratarem de testes como /Amostra pH/, /Coagulante ph/, /Remoção resultado ensaio/ e /Ensaio amido/ e outros que não têm representação lógica como /Congresso brasileiro/.

Texto 7: **Estudo de tratabilidade de água residuária sintética simulando despejo líquido de coquerias**

✓ Total de palavras no texto: 2.066 palavras.

Indexação automática

✓ Foram localizados 367 termos significativos pelo sistema.

1. Amônia livre	2. Cianeto livre
3. DQO total	4. Efeito processo
5. Processo nitrificação	6. Efluente final sistema
7. Idade lodo	8. Matéria orgânico
9. Tratamento biológico	10. Tratamento unidade

Quadro 14: Palavras-chave extraídas do texto 7

Indexação manual

1. Coqueria	2. Tratamento da água
3. Lodo ativado	4. Tratamento de esgoto
5. Poluição da água	

Quadro 15: Palavras-chave atribuídas ao texto 7

A indexação automática atribuiu ao texto 710 palavras-chave. Notou-se grande extensividade nas extrações do protótipo quanto às fases do processo descritas no texto como: /DQO total/, /Cianeto livre/, /Processo nitrificação/ e /Idade lodo/. E extraiu ainda, /Tratamento biológico/ e /Matéria orgânico/, podendo estes serem considerados conceitos mais gerais. Entretanto, não é representado o material que sofre o processo ou sua origem, no texto aplicado, tendo estes grande relevância por tratar-se de resíduos da indústria de metalurgia ou esgoto industrial.

A extração de /Tratamento unidade/ é observada, no caso, como uma palavra-chave fundamental na situação de pós-coordenação com as fases do processo de tratamento. /Efluente final sistema/ não tem nenhuma representação lógica dentro do conteúdo tratado.

Na indexação manual as palavras-chave demonstraram conceitos mais gerais como /Tratamento de Esgoto/ e /Poluição da água/.

/Coqueria/ é a fonte que origina o material a ser tratado, estando, portanto, intimamente ligado ao universo expresso pelo texto. /Lodo ativo/ como o meio a ser tratado também possui grande representatividade apesar de existir uma relação de ambigüidade entre esta atribuição e /Poluição da água.

/Tratamento da água/ foge ao contexto, talvez pelo texto tratar de despejos líquidos ou água residuária, havendo assim uma interpretação inadequada do conteúdo.

Neste texto aparece claro que a indexação automática está direcionada para um sistema que permita a pós-coordenação mais efetiva, com alto grau de

extensividade e a indexação manual trata de conceitos mais gerais sem muitas inter-relações entre as palavras-chave.

Texto 8: **Parados e sufocados**

✓ Total de palavras no texto: 1.364 palavras.

Indexação automática

✓ Foram localizados 181 termos significativos pelo sistema.

1. Meio ambiente	2. Região metropolitana

Quadro 16: Palavras-chave extraídas do texto 8

Indexação manual

1. Meio ambiente	3. Problemas respiratórios
2. Poluição do ar	4. Rodízio de automóveis

Quadro 17: Palavras-chave atribuídas ao texto 8

Nos dois métodos utilizados foi reconhecida a palavra-chave /Meio ambiente/. As demais vistas, / Região metropolitano/ pela indexação automática e / Problemas respiratórios/, /rodízio de automóveis/ atribuídos pela indexação manual, são produtos que podem ser considerados como termos relacionados a / Meio ambiente/.

Foi também atribuído /Poluição do ar/ pela indexação manual, que pode ser considerado como termo específico do campo semântico Saneamento básico, conceito geral, e/ou termo relacionado a /Meio ambiente/. Este é o texto em que as indexações demonstraram maior consistência quanto aos termos atribuídos. Não houve ocorrência de conceitos vazios ou não relacionados ao conteúdo expresso.

CONCLUSÃO

Pode-se constatar que se não fossem as ocorrências de sinonímias, os produtos automáticos aproximariam-se potencialmente dos manuais. Observamos que a indexação manual gerou léxicos documentários genéricos e representações que não possibilitariam uma recuperação pertinente do contexto.

Pudemos observar que a indexação automática está direcionada para um sistema que permita a pós-coordenação mais efetiva, com alto grau de extensividade e que a indexação manual tratou de conceitos mais gerais sem muitas inter-relações entre as palavras-chave.

A precisão é obtida não pela pequena quantidade de termos, mas pela efetiva representatividade do produto documentário, ou seja, possibilidade de remeter a determinados sistemas de significação dos textos.

No processo de indexação manual, mesmo tendo uma quantidade menor de termos com relação à automatizada, o significado dos léxicos teria uma abrangência grande. Nesse caso, também, a revocação seria alta por possibilitar a recuperação de documentos relevantes, mas não precisos.

Não temos a pretensão de avaliar o processo de indexação manual, entretanto, a partir dos resultados, as representações remetiam a vários sentidos possíveis. A análise que nos propusemos fazer não estava direcionada em averiguar se foram atribuídos os mesmos termos para ambos os processos: foi pautada no significado que os léxicos teriam, ou seja, representavam ou não os contextos dos discursos científicos. Com isso seria possível obter resultados precisos ou não em uma busca em sistema automatizado.

Os problemas apresentados na indexação manual podem ser atribuídos a formação dos indexadores. A aplicação e o conhecimento dos métodos de AD e das interfaces Lingüística e Terminologia permitiria impor rigor ao tratamento de informações e à construção de linguagens com fins documentários. Ao utilizarem listas alfabéticas estarão se assemelhando às representações automáticas, pois uma lista alfabética não garante significação, por operar a partir da palavra, com isso, remeter a vários sentidos possíveis. O parâmetro que norteia a construção de linguagens documentárias é o sistema: " (...) onde nada significa de forma isolada, mas a partir de cada palavra em relação às outras (...)" A estrutura significa a presença de dois termos vinculados por uma relação. Após a definição semântica

dos termos "(...) efetua-se uma relação dos termos, reconhece-se a polissemia e propõe-se uma interpretação unívoca do conceito/termo" (RODRIGUES, 1999, p. 1).

Observamos que o processo automático tenta remontar gramaticalmente um conceito, já que os termos compostos foram fatorados, mas aproxima-se mais de uma pré-coordenação sem o uso de sinais, como o hífen em ambas as áreas. E que o termo composto fatorado na forma sintática perde significado, por operar a partir da palavra. Na maior parte dos estudos sobre sistemas automáticos de indexação realiza-se a representação de uma informação sem a sua análise, ou seja, indexação assistida por computador, mas no caso estudado, a análise seria feita posteriormente para efetuar correções/ acréscimos no vocabulário alimentado para o protótipo.

De acordo com a análise, os resultados demonstraram uma maior ambigüidade na área de Humanas relativamente à de Exatas que demonstrou uma representação mais voltada às particularidades, portanto, gerando pouca ambigüidade. É possível afirmar que a representação automática pode vir a ser empregada na área de Exatas com uma margem de acertos considerável.

Foi possível averiguar que, em determinadas representações, a automática conseguiu remeter aos significados dos discursos, e a manual, em outras, necessitava ser precisa, não como referência a fins quantitativos, mas em aspectos semânticos. A automática aproximou-se da manual provavelmente por terem sido extraídos termos que remetiam a seus referentes: aos objetos de uma realidade extralinguística, entretanto, como já dissemos teria-se que efetuar uma análise após o processo automático, na tentativa de consistir os produtos documentários, quer acrescentando-os, quer corrigindo ou eliminando-os, em razão dos níveis altos de exaustividade, portanto seria uma indexação assistida por computador.

O método poderia ser melhor utilizado para uma efetiva indexação automática se reconhecesse termos compostos, relações semânticas e sistematizasse os métodos automáticos de análise., porque só é possível representar a partir do conhecimento, ou seja conhecer é condição necessária para representar e transferir informações. Conhecer no âmbito da AD é analisar e representar à partir de LD's estruturadas segundo parâmetros linguístico-terminológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARO, R.K.O.F. **Contribuição da análise do discurso para uma análise documentária: o caso da documentação jornalística.** São Paulo : ECA/USP, 1991. Dissertação (Mestrado em Ciência da Informação) - Escola de Comunicações e Artes, Universidade de São Paulo.

CANTARELLI, Elisa Maria Pivetta. **Acesso a base de dados através da linguagem natural**. [online]. Uruguai, 1998. [Trabalho de Conclusão de Curso - Universidade Regional Integrada do Alto Uruguai]. Disponível na internet: <http://www.biblio.virtual.urifw.tche.br/bvinf/tc1998/elisa.htm>

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

CASTILHO, Virgínia. **Para uma indexação automática:** métodos de análise de textos em sistemas informatizados aplicados à indexação. São Paulo, 1995. [Trabalho de Conclusão de Curso - Departamento de Biblioteconomia e Documentação, Escola de Comunicação e Artes, Universidade de São Paulo]

CINTRA, Ana Maria Marques. **Análise de texto/análise de discurso e possíveis relações com a análise documentária** : texto provisório 1. São Paulo : [s.n.]., 1994 Texto utilizado na disciplina Lingüística e Documentação do Departamento de Biblioteconomia e Documentação da Escola de Comunicação e Artes da Universidade de São Paulo.

_____ . **Análise de texto/análise de discurso e possíveis relações com a análise documentária** : texto provisório 2. São Paulo : [s.n.]., 1994 Texto utilizado na disciplina Lingüística e Documentação do Departamento de Biblioteconomia e Documentação da Escola de Comunicação e Artes da Universidade de São Paulo.

COPI, I.M. **Introdução à lógica**. São Paulo : Mestre Jou, 1978.

COULON, Daniel, KAYSER, Daniel. **Informática e linguagem natural**: uma visão geral dos métodos de interpretação de textos escritos. Brasília : IBICT, 1992.

COYAUD, M. Étude théorique des différentes méthodes d'analyse automatique des documents. In: COYAUD, M., SIOT-DECAUVILLE, N. **L'analyse automatique des documents**. Paris : Mouton. p.2-57, 1967.

CUNHA, Izabel Maria Ribeiro Ferin. **Do mito à análise documentária**. São Paulo: Edusp, 1990. (Teses, 11).

DIAS, Cláudia Augusto. Terminologia: conceitos e aplicações. **Ciência da Informação**. Brasília, v.29, n.1, p. 90-92, jan./abr. 2000.

ECO, Humberto. **O signo**. Lisboa : Presença, 1997.

FERNEDA, Edberto. **Construção automática de um thesaurus retangular**. Campina Grande, Paraíba, 1997. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal da Paraíba, 1997.

GARCÍA GUTIÉRREZ, A. **Analisis documental del discurso periodístico**. Madrid : CTD, Centro de Tratamiento de la Documentación, 1992.

GARDIN, J.-C. **La logique du pausible**. Paris : Maison des sciences de l'homme, 1981.

GOMES, H.E. O indexador face às novas tecnologias de informação. **Trans-in-formação**, Campinas, v. 1, n.2, p. 161-171, maio/ago. 1989.

GOMES, Henriette Ferreira. O ambiente informacional e suas tecnologias na construção dos sentidos e significados. **Ciência da Informação**, Brasília, v.29, n.1, p.61-70, jan./abr. 2000.

HERMANS, A. La définition des termes scientifiques. **Meta**, v.34, n.3, p. 529-532.

ISO 704. **Princípios e métodos da atividade terminológica**, 1994. Proposta da norma brasileira.

ISO 1087. **Terminologia – vocabulário**, 1994. Proposta da norma brasileira [trad. e adapt. Grupo ABNT/IBICT]

KOBASHI, Nair Yumiko. **Análise documentária**: metodologias para indexação e resumo.[s.l] : [s.n.], 1995.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação: os sintagmas nominais. **Ciência da informação online**, Brasília, v.25, n.2, p. 1-17, 1996. Disponível na internet: <http://www.ibict.br/cionline/250296/25029605.htm>

LANCASTER, F.W. **Indexação e resumos:** teoria e prática. Briquet de Lemos/Livros. 1993.

LARA, Marilda Lopes Ginez de. Linguagens documentárias: instrumentos de mediação e comunicação. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v.26, n.1/2, p. 72-80, jan./jun. 1993.

_____ . **Representação e linguagens documentárias**: bases teórico-metodológicas. São Paulo : M.L.G. Lara, 1999. Tese (Doutorado em Ciência da Comunicação) - Escola de Comunicação e Artes, Universidade de São Paulo.

LIMA, Vânia M.A. **Comunicação e representação documentária**. São Paulo : APB, 1999. (Ensaios APB, 62).

MAMFRIM, Flávia Pereira Braga. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ciência da Informação**, Brasília, v.20, n. 2, p. 191-203, jul./dez. 1991.

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

NOVELLINO, Maria Salet Ferreira. A linguagem como meio de representação ou de comunicação da informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.3, n.1, p. 137-146, jul./dez. 1998

ROBREDO, Jaime. A indexação automática de textos: o presente já entrou no futuro. In: MACHADO, U.D. **Estudos avançados de Biblioteconomia e Ciência da Informação**. Brasília : ABDF, 1982.

_____ . Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, Brasília, v.20, n. 2, p.130-136, jul/dez. 1991.

RODRIGUES, Willame Santos. **Auxílio da lingüística na construção de linguagens documentárias**: conceitos utilizados no arranjo dos termos. São Paulo, 1999. [Trabalho apresentado à disciplina Linguagem de Indexação II – Faculdade de Biblioteconomia e Ciência da Informação].

ROLE, François. De la lettre au sens: les recherche en texte integral. **Documentaliste**, v.30, n. 3, p. 140-146, 1993.

TÁLAMO, Maria de Fátima Gonçalves Moreira. **Linguagem documentária.** São Paulo: APB, 1997. (Ensaios APB ; 45)