UNIVERSIDADE ESTADUAL PAULISTA
CAMPUS DE MARÍLIA
Faculdade de Filosofia e Ciências

CRISTINA MIYUKI NARUKAWA

ESTUDO DE VOCABULÁRIO CONTROLADO NA INDEXAÇÃO AUTOMÁTICA: APLICAÇÃO NO PROCESSO DE INDEXAÇÃO DO SISTEMA DE INDIZACIÓN SEMIAUTOMATICA (SISA)

Marília/2011

CRISTINA MIYUKI NARUKAWA

ESTUDO DE VOCABULÁRIO CONTROLADO NA INDEXAÇÃO AUTOMÁTICA: APLICAÇÃO NO PROCESSO DE INDEXAÇÃO DO SISTEMA DE INDIZACIÓN SEMIAUTOMATICA (SISA)

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Campus de Marília, como requisito parcial para obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Produção e Organização da Informação

Orientadora: Dra. Mariângela Spotti Lopes Fujita (Universidade Estadual Paulista/Marília)

Co-orientador: Dr. Isidoro Gil Leiva (Universidade de Murcia/Espanha)

Apoio: Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP)

Marília/2011

Ficha Catalográfica
Serviço de Biblioteca e Documentação – UNESP - Campus de Marília

Narukawa, Cristina Miyuki.
N237e Estudo de vocabulário controlado na indexação automática : aplicação no processo de indexação do Sistema de Indización SemiAutomatica (SISA) / Cristina Miyuki Narukawa. – Marília, 2011.
f. ; 30 cm.

Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2011.
Orientador: Profa. Dra. Mariângela Spotti Lopes Fujita.
Co-orientador: Prof. Dr. Isidoro Gil Leiva

1. Indexação automática. 2. Indexação – Software. 3. Sistema de recuperação da informação. 4. Sistema de Indización SemiAutomatica. 5. Linguagem de indexação. 6. Vocabulário controlado. I. Autor. II. Título.

CDD 029.5

CRISTINA MIYUKI NARUKAWA

ESTUDO DE VOCABULÁRIO CONTROLADO NA INDEXAÇÃO AUTOMÁTICA: APLICAÇÃO NO PROCESSO DE INDEXAÇÃO DO SISTEMA DE INDIZACIÓN SEMIAUTOMATICA (SISA)

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Campus de Marília, como requisito parcial para obtenção do título de Mestre em Ciência da Informação.
Área de concentração: Informação, Tecnologia e Conhecimento
Linha de Pesquisa: Produção e Organização da Informação

Banca Examinadora

_____
Dra. Mariângela Spotti Lopes Fujita
Faculdade de Filosofia e Ciências da Universidade Estadual Paulista (UNESP)
Campus de Marília

_____
Dr. Renato Rocha Souza
Escola de Matemática Aplicada da Fundação Getúlio Vargas (FGV) – Rio de Janeiro

_____
Dr. José Augusto Chaves Guimarães
Faculdade de Filosofia e Ciências da Universidade Estadual Paulista (UNESP)
Campus de Marília

Marília, 22 de junho de 2011.

Dedico a minha família e aos meus amigos por todo apoio, carinho, incentivo e por significarem tudo para mim

AGRADECIMENTOS

A *Deus* por me fortalecer nos momentos difíceis mostrando que podemos superá-los e ampliar as oportunidades.

À orientadora *Professora Mariângela* por todo apoio, profissionalismo nas orientações, paciência e, sobretudo pela confiança que depositou nesta parceria de pesquisa.

Ao co-orientador *Professor Isidoro* por estar sempre presente, esclarecendo prontamente as minhas dúvidas e por ampliar minha visão sobre a pesquisa.

Ao *Professor Renato Rocha Souza* e ao *Professor José Augusto Chaves Guimarães* por todas as sugestões para avanço da pesquisa apresentadas durante o exame de qualificação e defesa da dissertação.

A minha família, em especial ao meu pai *Yoshio*, minha mãe *Tokiko* e meus irmãos *Clara*, *Marta*, *Aline*, *Mutsuki* e *Megumi* por tudo que sempre fizeram por mim.

A *Tiemi*, *Rosangela*, *Alessandra*, *Aline, Andressa, Camila*, *Mariana Inácio, Cibele* e *Bruno, Stela, Maria dos Remédios, Carlos Augusto,Vera, Fabiana* por todo apoio, conversas, troca de experiências e momentos que vivenciamos.

Aos c*olegas de trabalho* do Serviço de Biblioteca e Documentação da Faculdade de Direito da Universidade de São Paulo (USP).

À *Cassia Gatti* da Coordenadoria Geral de Bibliotecas da UNESP.

Aos *colegas, professores e funcionários* do Programa de Pós-graduação em Ciência da Informação da UNESP/Campus de Marília

Aos *funcionários* do Escritório de Pesquisa da UNESP/Campus de Marília

Aos *funcionários* da Biblioteca da Faculdade de Filosofia e Ciências da UNESP/Campus de Marília

A Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento da pesquisa.

A Biblioteca Nacional de Agricultura (BINAGRI) pela autorização de uso do vocabulário controlado ThesAgro para o experimento da pesquisa.

...Enfim, esta pesquisa é fruto da contribuição de várias pessoas que estiveram comigo compartilhando conhecimentos, experiências e sentimentos que de alguma forma, estão registrados em minhas palavras. Por isso, gostaria de deixar aqui, minha mais profunda gratidão a todos que contribuíram com mais uma realização.

[...] a pedra angular da construção informacional é
a análise da informação, a representação e organização do conhecimento.
As tecnologias ajudam a aprimorar o processo, e os dois juntos –
processo informacional e tecnologia – permitem sentar as bases de
novas aplicações que abrem novos horizontes [...]

No fim das contas, se o sistema funciona
não é por obra e graça de uma máquina,
mas daqueles que o imaginaram, o construíram,
o alimentaram e o fizeram funcionar
Jaime Robredo (2005)

RESUMO

A indexação automática é um processo complexo e delegar a atividade de atribuição de termos aos sistemas automáticos requer análise, tanto dos métodos, quanto das características dos instrumentos de indexação. Desse modo, propomos investigar a atuação de vocabulário controlado neste processo a partir da análise dos resultados de aplicação do vocabulário ThesAgro no Sistema de Indización SemiAutomatica (SISA), com objetivos de identificar as características que definem e distinguem os tipos de vocabulários; analisar propostas metodológicas e sistemas de indexação; aplicar o ThesAgro no sistema SISA em análise comparativa com a indexação manual da Biblioteca Nacional de Agricultura (BINAGRI), e analisar os fatores intervenientes que apontam os problemas ocasionados à indexação automática. De modo geral, buscamos contribuir com o desenvolvimento do tema ao levantar subsídios para adaptação de vocabulários controlados. Realizamos uma revisão teórica sobre sistemas de indexação automática e um experimento aplicando o ThesAgro no sistema SISA com 100 artigos da área agrícola, especificamente sobre fruticultura. Utilizamos, como parâmetro de avaliação, a indexação manual realizada pela BINAGRI e análise comparativa com os resultados de pesquisa anterior em que se avaliou o desempenho do vocabulário Descritores em Ciências da Saúde (DeCS) no referido sistema. A partir da análise dos resultados constatamos que o vocabulário condiciona os resultados do processo de indexação automática e, portanto, é necessário compreendê-lo, considerando os métodos de identificação das unidades representativas da informação, aplicação de tratamento linguístico, características da área do conhecimento, relações semânticas, idioma, atualização, uso de vocabulários em conjunto e sua relação com outros instrumentos de indexação automática, uma vez que os sistemas automáticos ainda não têm proporcionado resultados tão favoráveis na identificação dos conteúdos informacionais dos documentos, o que permite concluir que é preciso, antes de tudo, analisar as condições exigidas para possibilitar uma indexação de qualidade.

Palavras-chave: Indexação Automática. Vocabulário Controlado. Sistemas de Indexação Automática. Sistema de Indización SemiAutomatica (SISA). Avaliação da indexação. Linguagem de Indexação.

ABSTRACT

Automatic indexing is a complex process, and delegating the attribution of terms to automatic systems requires analyzing not only the methods, but also the features of indexing instruments. Thereby, we propose to investigate the role of controlled vocabulary in such process, based on the analysis of results from the application of ThesAgro vocabulary in the Semi-Automatic Indexing System (SISA - *Sistema de Indización SemiAutomatica* -), with the purposes of identifying the characteristics which define and distinguish the types of vocabularies; analyzing methodological proposals and indexing systems; applying the ThesAgro in the SISA, making a comparative analysis related to the manual indexing by the National Library of Agriculture (BINAGRI - *Biblioteca Nacional de Agricultura*), and analyzing the intervening factors pointing to the occurrence of problems concerning automatic indexing. As a general matter, we seek to contribute to the development of this theme by raising subsidies for adapting controlled vocabularies. We have performed a theoretical review on automatic indexing systems, and an experiment applying the ThesAgro in the SISA, with 100 articles on agriculture , specifically about fruit production. The manual indexing performed by BINAGRI and the comparative analysis with the results from a previous research, which evaluated the performance of the vocabulary from the Health Sciences Descriptors (DeCS - *Descritores em Ciências da Saúde*) in the before mentioned system, have served as the evaluation parameter. The analysis of results allows us to conclude that the vocabulary conditions the results of the automatic indexing process. Thus, it is necessary to understand it, considering the identification methods of the information representative units, application of linguistic treatment, features of the knowledge field, semantic relations, idiom, updating, use of combined vocabularies and its relation to other automatic indexing instruments, as the automatic systems have not yet yielded much favorable results in identifying the informational content in documents. This leads to the conclusion that it is essential, above all, to analyze the required conditions which enable a quality indexing.

Keywords: Automatic indexing. Controlled Vocabulary. Automatic Indexing Systems. *Sistema de Indización Semiautomático* (SISA). Indexing Evaluation. Indexing Language.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AGLINET	-	Rede Internacional de Bibliotecas Agrícolas
Agris	-	Sistema Internacional de Informação para a Ciência e Tecnologia Agrícola
AGROBASE	-	Base Bibliográfica da Agricultura Brasileira
BDBINAGRI	-	Base de dados BINAGRI
BDBIREME	-	Base de dados BIREME
BDSISA	-	Base de dados SISA
BDTD	-	Biblioteca Digital Brasileira de Teses e Dissertações
BINAGRI	-	Biblioteca Nacional de Agricultura
BINDEX	-	Bilingual Automatic Parallel Indexing and Classification
BIREME	-	Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde
CADIS	-	Computer Aided Document Indexing System
CERN	-	Laboratório Europeu de Física de Partículas
CETIS	-	Européen pour le Traitement de l'Information Scientifique
CIAT	-	International Center for Tropical Agriculture
CR	-	Começo do Resumo
CRG	-	Classification Research Group
CTE	-	Começo do Texto
CTI	-	Começo do Título
DeCS	-	Descritores em Ciências da Saúde
DESY	-	Deutsche Elektronen-Sybchrotron
Eurovoc	-	Thesaurus Multilingue da União Europeia
FAIRS	-	Fully Automatic Information Storage and Retrieval
FAO	-	Food and Agriculture Organization of the United Nations
FR	-	Fim do Resumo
FTE	-	Fim do Texto
FTI	-	Fim do Título
IDF	-	Inverse Document Frequency Weight
IICA	-	Instituto Interamericano de Cooperação para a Agricultura
KWAC	-	Key Word And Context
KWAT	-	Key Word And Title
KWIC	-	Keyword In Context
KWIT	-	Key Word In Title
KWOC	-	Key Word Out of Context
KWOT	-	Key Word Out of Title
LCSH	-	Library of Congress Subject Headings
LIPHIS	-	Linked Phrase Indexing System
LISA	-	Library Information Science Abstract
LISTA	-	Library, Information Science and Technology Abstracts
LO	-	Leady Only
MAPA	-	Ministério da Agricultura, Pecuária e Abastecimento

MCE	-	Multimedia Concept Extraction
Mesh	-	Medical Subject Headings
MGI	-	Módulo Gerador de Índice
MGO	-	Módulo Gerador de Ontologias
MySQL	-	Structured Query Language
NEPHIS	-	Nested Phrase Indexing System
NLM	-	National Library of Medicine
PDF	-	Portable Document Format
PHL	-	Personal Home Library
PLN	-	Processamento de Linguagem Natural
POPSI	-	POstulated-based Permuted Subject Indexing Language
PRECIS	-	PREserved Context Indexing System
SGBD	-	Sistema de Gerenciamento de Banco de Dados
SIDALC/AGRI2000	-	Sistema de Información y Documentación Agropecuário de America
SiRILiCO	-	Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia
SISA	-	Sistema de Indización SemiAutomatica
SMA	-	Sub-módulo Atomizador
SMART	-	Sistem for the Manipulation and Retrieval of Text
SMOB	-	Sub-módulo de Ontologia Básica
SMOF	-	Sub-módulo de Ontologia Formada
SMOSe	-	Sub-módulo Semântico
SMOSi	-	Sub-módulo Sintático
SN	-	Sintagma Nominal
SNIDA	-	Sistema Nacional de Informação Agrícola
TE	-	Termo Específico
TG	-	Termo Geral
ThesAgro	-	Thesaurus Agrícola Nacional
TICs	-	Tecnologias de Informação e Comunicação
TR	-	Termo Relacionado
UNESCO	-	Organização das Nações Unidades para a Educação, a Ciência e a Cultura
UNESP	-	Universidade Estadual Paulista “Júlio de Mesquita Filho”
UNISIST	-	World Information System for Science and Technology
UP	-	Usado Para
USDA/NAL	-	United State Department of Agriculture
USE	-	Use
UTC	-	Unidades Terminológicas Complexas
W3C	-	World Wide Web Consortium
XML	-	Extensible Markup Language

SUMÁRIO

INTRODUÇÃO

A temática abordada nesta pesquisa está circunscrita ao campo da Ciência da Informação, especificamente na linha de pesquisa "Produção e Organização da Informação".

A Ciência da Informação, como campo científico, dedica-se à prática profissional e às questões teórico-metodológicas relacionadas aos problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação, considerando as vantagens advindas das modernas tecnologias informacionais para tratar dessas questões (SARACEVIC, 1996).

Como explica Saracevic (1996), a Ciência da Informação originou-se no contexto da revolução científica e técnica do período Pós-Segunda Guerra Mundial, marcada pela *explosão informacional* — crescimento da produção técnica e científica —, fato que explica a necessidade de se pensar em alternativas para organizar a informação e permitir seu uso eficaz. Bush (1945) já indicava em seu artigo "*As we may think*" esse problema. Assim, propôs soluções tecnológicas de organização fundamentadas em associações de ideias, simulando a mente humana — ao contrário das soluções propostas até então, que Bush considerou superficiais e inflexíveis.

No entanto, é necessário lembrar que a preocupação com o acesso à crescente produção científica e técnica já era considerada por bibliotecários e documentalistas, que propunham formas de organizar a informação para permitir seu acesso dentro das possibilidades tecnológicas daquela época. Podem ser citados, como exemplo, o sistema de classificação conhecido como "Classificação Decimal Universal" (CDU), criado por Paul Otlet e Henry La Fontaine; a concepção de indexação sistemática proposta por Kaiser; as listas de cabeçalhos de assunto e a *Colon Classification* de Ranganathan. Estas formas de organização, advindas da tradição biblioteconômica e da documentação, foram criadas entre o final do séc. XIX e o início do séc. XX e hoje oferecem princípios teóricos importantes para a área de Ciência da Informação.

Ainda que essas discussões remontem ao séc. XIX, os problemas ainda permanecem e parecem ainda mais intensos quando se analisa a dificuldade em recuperar informações disponíveis no ambiente caótico da *Web*.

A necessidade de tornar a informação disponível e, mais do que isso, acessível, e de permitir o seu uso pelas pessoas, deve considerar todas as questões que perpassam os processos de produção, coleta, tratamento ou organização, recuperação, disseminação e uso da

informação, entendendo-se que, a partir desse uso, um novo conhecimento pode ser gerado, propiciando uma nova produção de informação (GUIMARÃES, 2003).

Nesse contexto, a análise automática de textos é considerada uma área de pesquisa importante há mais de trinta anos e continua sendo de grande interesse na Ciência da Informação. Atualmente, existe a preocupação em oferecer acesso rápido à literatura técnico-científica e é possível utilizar o computador no processamento de dados e informações (ROBREDO, 2005). De acordo com Robredo (2005, p. 170) a indexação automática consiste em "qualquer procedimento que permita identificar e selecionar os termos que representam o conteúdo dos documentos, sem a intervenção direta do indexador".

A aplicação da indexação automática desenvolveu-se como alternativa na análise e representação da informação diante do crescimento exponencial de documentos. Robredo (2005) afirma que a necessidade de indexar grandes volumes de informação, em um tempo curto para manter as bases de dados atualizadas, tornou inviável pensar na indexação manual (humana ou intelectual) como única forma de analisar e codificar o conteúdo dos documentos. Dessa forma, Robredo (2005) compreende que as pesquisas relacionadas à indexação automática devem-se desenvolver paralelamente às pesquisas sobre indexação manual (humana ou intelectual).

De acordo com Gil Leiva (1999), o desenvolvimento da indexação automática ocorreu em três momentos históricos: o dos métodos estatísticos, o dos métodos linguísticos e o dos métodos mistos. Como explica Gil Leiva (1999), os métodos estatísticos foram os primeiros a surgir e tiveram influência das ideias de Zipf, que, em 1949, enunciou o princípio do mínimo esforço — segundo o qual a relação entre a frequência das palavras e a posição que essas ocupam na ordem frequencial tem valor constante. A partir dessas ideias, Luhn sugeriu, em 1957, que a frequência das palavras em um texto está ligada à utilidade que teriam na indexação, e, a partir da ideia de frequência das palavras como critério para indexação, outras concepções começaram a surgir.

Os métodos linguísticos começam a ser utilizados na indexação automática a partir do uso de computadores para compreender a linguagem, passando por experimentos de tratamento sintático nos anos 1970 e, a partir dos anos 1980, há a ampliação dos estudos de processamento em linguagem natural ao nível semântico (GIL LEIVA, 1999).

Como exceção aos primeiros sistemas com métodos puramente estatísticos, verificam-se os métodos mistos ou híbridos de indexação automática que reúnem aportes da estatística, da linguística textual e, ainda, utilizam tesauros como instrumento de controle de vocabulário,

contribuindo para eliminar problemas como os causados pela sinonímia e pela não identificação das funções sintáticas dos termos, proporcionando benefícios à revocação na recuperação da informação (GIL LEIVA, 1999; GUIMARÃES, 2000).

Com relação ao modo com que os termos de indexação são selecionados, Lancaster (2004) distingue dois tipos de indexação: a indexação por extração automática e a indexação por atribuição automática. Na indexação por extração automática, as palavras ou expressões que aparecem no texto são extraídas e utilizadas para representar o texto como um todo; ou seja: a indexação é realizada a partir da linguagem natural. Por sua vez, a indexação por atribuição automática consiste no processo de representação do conteúdo do documento mediante termos selecionados de algum vocabulário controlado. Esse processo, segundo Lancaster (2004), é realizado também na indexação por seres humanos, mas é considerado difícil quando aplicado a computadores.

Para efetuar a indexação por atribuição automática é necessário "[...] desenvolver, para cada termo a ser atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer frequentemente nos documentos [...]" (LANCASTER, 2004, p. 289). Metodologia nesse sentido é utilizada no *Sistema de Indización SemiAutomatica* (SISA), no qual se efetua a indexação por comparação entre o documento — constituído por título, resumo e texto — e um vocabulário controlado, partindo de critérios de frequência preestabelecidos pelo sistema para propor os termos de indexação.

O vocabulário controlado é "essencialmente uma lista de termos autorizados" (LANCASTER, 2004, p. 19), que constitui uma linguagem de indexação. As linguagens de indexação são entendidas por Cintra *et al*. (2002, p. 33) como linguagens "[...] construídas para indexação, armazenamento e recuperação da informação e correspondem a sistema de símbolos destinados a 'traduzir' os conteúdos dos documentos". Nesse sentido, são considerados instrumentos intermediários, por meio dos quais se realiza a tradução das informações que foram identificadas e selecionadas na análise do conteúdo do documento para representação. Em um segundo momento, servem para traduzir as necessidades informacionais do usuário em termos de busca para recuperação. É quando ocorre a compatibilidade entre a representação dessa necessidade de busca e a representação do conteúdo temático dos documentos que, efetivamente, ocorre a recuperação da informação.

Portanto, no processo de indexação automática, os vocabulários controlados atuam no próprio processo de análise automática do documento e na representação, ou seja, condicionam os resultados na atribuição de descritores. Por isso há a necessidade de

considerar todos os seus atributos, além da necessidade de considerar a sua atuação associada aos métodos de indexação automática pelos quais se realiza a análise do conteúdo temático dos documentos.

Em pesquisa realizada por Narukawa, Gil Leiva e Fujita (2009), foi proposta a investigação da indexação com ênfase no processo da indexação automática por meio de análise dos índices de consistência na indexação bem como de exaustividade e precisão na recuperação da informação. Para isso, realizou-se análise comparativa entre a indexação automática com o uso da linguagem Descritores em Ciências da Saúde (DeCS) e a indexação manual do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME), buscando contribuir com o aperfeiçoamento do SISA e com o desenvolvimento teórico da indexação automática.

Os resultados da pesquisa mostraram a necessidade de adaptação do vocabulário controlado, principalmente no que diz respeito ao estabelecimento de relações entre os termos, uma vez que as relações semânticas entre conceitos devem ser mantidas durante a indexação automática para garantir qualidade na recuperação da informação. Deve ser considerado também que, na indexação manual, o método de atribuição de descritores é baseado na investigação dos conceitos, havendo reflexão sobre os assuntos apresentados no artigo — reflexão esta que caracteriza o processo cognitivo humano na atribuição de significado durante a compreensão. Por outro lado, na indexação automática do SISA verifica-se um processo diferente, ou seja, trata-se de um processo automático baseado na frequência de palavras e que, portanto, responde a regras determinadas, o que tem influência no emprego do vocabulário controlado.

Desse modo, verificamos que a **problemática** subjacente à dissertação está na necessidade de investigar os aspectos relacionados aos vocabulários controlados aplicados em indexação automática, considerando a complexidade do processo automático e as características dos tradicionais instrumentos de representação da informação.

Na elaboração de vocabulários controlados para indexação automática é necessário considerar os critérios de análise de conteúdo adotados pelo sistema automático para prever as possíveis variações da linguagem natural, ou seja, um vocabulário controlado que considere as características particulares da indexação automática.

Dessa forma, **propomos** investigar a atuação de um vocabulário controlado no processo de indexação automática a partir da análise dos resultados de aplicação do

vocabulário controlado Thesaurus Agrícola Nacional (ThesAgro) no sistema SISA em artigos de periódicos científicos.

A partir dessa proposta, temos o **objetivo geral** de contribuir com o desenvolvimento da indexação automática, investigando a adaptação de vocabulários controlados para serem aplicados no processo de indexação automática. Os **objetivos específicos** são:

(1) Identificar e analisar as principais características que definem e distinguem os tipos de vocabulários controlados;

(2) Analisar as possibilidades para indexação automática oferecidas pelas propostas metodológicas e pelos sistemas de indexação automática, buscando verificar os aspectos relacionados à aplicação de vocabulários controlados;

(3) Aplicar o vocabulário controlado ThesAgro no processo de indexação automática do SISA, em análise comparativa com a indexação manual realizada pela Biblioteca Nacional de Agricultura (BINAGRI);

(4) Analisar os fatores intervenientes na atuação dos vocabulários controlados ThesAgro e DeCS na indexação automática do SISA e verificar, sob uma perspectiva mais ampla, os problemas enfrentados na indexação automática.

As mudanças que ocorrem na sociedade são, muitas vezes, impostas por influências políticas, econômicas, sociais, culturais e tecnológicas que caracterizam uma época, oferecendo novas alternativas para propor soluções. Entretanto, entendemos que essas influências não devem ser incorporadas passivamente como inovações sem uma análise crítica das possíveis interferências que podem ocasionar.

Essa análise crítica também deve ser considerada no momento de desenvolver e avaliar sistemas de organização da informação, principalmente porque as metodologias exploradas para organizar informação refletem um modo de pensar e interferem no próprio desenvolvimento do conhecimento de uma sociedade. As tecnologias de informação e comunicação influenciaram os ambientes profissionais e acabam por exigir, ao mesmo tempo, reflexões teóricas que possam fundamentar essas práticas.

Nesse sentido, **justifica-se** a análise de propostas que se originaram na indexação automática considerando as contribuições que podem oferecer, as suas limitações e as suas possibilidades de aperfeiçoamento, no que se refere à aplicação de vocabulários controlados. Pressupõe-se que é importante reconhecer e considerar os princípios subjacentes aos

vocabulários controlados na aplicação em sistemas de indexação automática para que pesquisadores e profissionais da informação possam avaliar criticamente as possibilidades de aplicações automáticas.

Como forma de melhor esquematizar o foco desta pesquisa, apresentamos a sistematização do problema, a nossa proposta, o objetivo geral da pesquisa e os objetivos específicos de cada capítulo:

QUADRO 1 - Relação entre os objetivos e os capítulos da pesquisa

	SISTEMATIZAÇÃO DA PESQUISA
Estrutura	**Delimitação**
Problema	Necessidade de investigar os aspectos relacionados aos vocabulários controlados aplicados em indexação automática, considerando a complexidade do processo e as características dos tradicionais instrumentos de representação da informação.
Proposta	Investigar a atuação de vocabulários controlados no processo de indexação automática a partir da análise dos resultados de aplicação do vocabulário controlado ThesAgro no sistema SISA em artigos de periódicos científicos.
Objetivo Geral	Contribuir com o desenvolvimento da indexação automática ao investigar a adaptação de vocabulários controlados para aplicação no processo de indexação automática.
Capítulo 2	**Objetivo específico 1:** Identificar e analisar as principais características que definem e distinguem os tipos de vocabulários controlados. **Título:** Os vocabulários controlados como linguagens de indexação
Capítulo 3 **Capítulo 5**	**Objetivo específico 2:** Analisar as possibilidades para indexação automática oferecidas pelas propostas metodológicas e pelos sistemas de indexação automática, buscando verificar os aspectos relacionados à aplicação de vocabulários controlados. **Título:** Indexação automática **Título:** Sistemas de indexação automática
Capítulo 4	**Metodologia** **Título:** Procedimentos metodológicos
Capítulo 6	**Objetivo específico 3:** Aplicar o vocabulário controlado ThesAgro no processo de indexação automática do SISA em análise comparativa com a indexação manual realizada pela BINAGRI. **Título:** Vocabulário controlado na indexação automática do SISA
Capítulo 7 **Considerações finais**	**Objetivo específico 4:** Analisar os fatores intervenientes na atuação dos vocabulários controlados ThesAgro e DeCS na indexação automática do SISA e verificar, sob uma perspectiva mais ampla, os problemas enfrentados na indexação automática. **Título:** Implicações sobre o uso dos vocabulários controlados no processo de indexação automática **Título:** Considerações finais

Fonte: Elaborado pela autora

Para atingir os objetivos propostos, desenvolvemos o experimento de aplicação do vocabulário controlado ThesAgro no sistema SISA com 100 artigos da área agrícola, especificamente sobre fruticultura, utilizando como parâmetro de qualidade para avaliação a indexação manual realizada pela BINAGRI. A análise dos dados obtidos em experimento foi fundamentada por pesquisa exploratória que tratou dos aspectos relacionados aos vocabulários controlados, à indexação automática e às propostas de sistemas de indexação automática.

A pesquisa é constituída por seis capítulos, além desta introdução. No capítulo 2, expomos o desenvolvimento e as características dos vocabulários controlados, em especial dos cabeçalhos de assuntos e dos tesauros, incluindo uma breve exposição sobre as ontologias (que se justifica em razão do seu emprego em alguns sistemas de indexação automática). No capítulo 3, abordamos os aspectos conceituais, históricos e metodológicos da indexação automática. Em seguida, no capítulo 4 tratamos dos procedimentos metodológicos da pesquisa, incluindo a revisão teórica, a aplicação do ThesAgro no SISA e a avaliação da indexação. No capítulo 5 são apresentadas e sistematizadas as principais características de vinte propostas metodológicas e sistemas de indexação automática. No capítulo 6 apresentamos os resultados obtidos a partir do experimento empregando o ThesAgro no sistema SISA. No capítulo 7, discutimos os resultados do experimento desta pesquisa comparando-os aos resultados do experimento da pesquisa anterior com uso do DeCS no sistema SISA e, finalmente, apresentamos os aspectos que nos dão indício de que, se forem mais bem analisados, poderão contribuir com o alcance de qualidade nos resultados proporcionados por indexação automática. E por fim, apresentamos as considerações finais, que apontam algumas sugestões de pesquisa.

2 OS VOCABULÁRIOS CONTROLADOS COMO LINGUAGENS DE INDEXAÇÃO

Neste capítulo, apresentamos um panorama do desenvolvimento dos vocabulários controlados, instrumentos aplicados há mais de um século para a representação da informação e que, ao longo dos anos, sofreram mudanças para se adaptarem às exigências de cada momento. Buscamos traçar e distinguir os aspectos dos vocabulários controlados e apresentar, mesmo que sucintamente, o surgimento das ontologias[1], modelos de representação da informação que têm recebido destaque na literatura de Ciência da Computação e, mais recentemente, na Ciência da Informação, por sua aplicabilidade em sistemas automáticos.

Ao longo do tempo, a sociedade produziu e acumulou conhecimento, em decorrência de suas reflexões, passando este à condição de informação à medida que foi registrado. A informação é insumo potencial para gerar novos conhecimentos, lançando, portanto, constantes mudanças em nossa sociedade. No entanto, para que esse conhecimento possa ser compartilhado e socializado, é necessário refletir sobre formas possíveis de organização da informação. Garantir sua disponibilização é, de certa forma, potencializar a apropriação dessa informação, permitindo que, assim, as pessoas a utilizem em prol das suas necessidades.

A linguagem tem papel fundamental, na medida em que permite ao homem expressar seus pensamentos, suas ideias e suas emoções. Configura-se como meio de expressão, preservação e socialização, tornando público todo conhecimento que foi também acumulado por meio da linguagem. Nas palavras de Vizcaya Alonso (1997), a linguagem, fruto da capacidade do homem de raciocinar através das ideias, é considerada o suporte material do pensamento, sendo um fenômeno social importante ao longo de toda a história humana.

A linguagem é uma forma de organização de informação, por meio da qual buscamos categorizar a realidade — realidade esta que pode ser categorizada de distintas maneiras (FIORIN, 2003). O que se busca é encontrar uma ordem para as coisas, já que um mundo caótico seria incompreensível, insuportável; por isso, o homem busca encontrar, em meio ao caos aparente, uma ordem, mesmo que subjacente, uma estrutura capaz de explicar as coisas (CINTRA *et al*., 2002).

Considerada um recurso de representação da realidade, a linguagem é incompleta em alguns aspectos, ambígua e, por vezes, limitada, acarretando implicações na comunicação.

[1] Apresentamos a ontologia por ser um instrumento capaz de auxiliar no processo de representação e recuperação da informação e, também, por sua aplicação em alguns sistemas de indexação automática, o que será verificado no capítulo 5.

Como tem afirmado Currás (2010), a linguagem classifica a realidade, limitando-a às nossas habilidades ou atitudes. De fato, tem-se constatado que os seres humanos elaboram mais ideias do que palavras para expressá-las.

No âmbito da Ciência da Informação, a linguagem apresenta valor inestimável, especialmente no processo de indexação e recuperação da informação. A linguagem expressa as ideias propostas pelo(s) autor(es) de uma obra, ou seja, é o meio de organização dos seus pensamentos para divulgação. É, ainda, a expressão da representação das ideias apresentadas na obra, como ocorre com a aplicação de vocabulários controlados na indexação e recuperação. Amplia o acesso às ideias do autor, servindo como elo entre a informação criada pelo autor e a necessidade dos indivíduos que potencialmente necessitam daquela informação.

Os vocabulários controlados, incluindo os cabeçalhos de assuntos e os tesauros, são um tipo de linguagem de indexação na qual a terminologia está controlada (LANCASTER, 2002). Linguagens de indexação[2] são linguagens construídas com o propósito de servir como instrumento de representação temática da informação. Pressupõe-se que, em um processo de análise, conceitos representativos do documento são identificados e selecionados. Em seguida, os conceitos são "traduzidos" nos termos dessa linguagem com o objetivo de representarem e tornarem-se pontos de acesso entre a informação do documento e aqueles que buscam essa informação.

Nesse contexto, o controle de vocabulário busca facilitar a representação consistente dos assuntos, atribuídos por indexadores e utilizados por usuários na recuperação, evitando a dispersão de informações relacionadas. Procura, ainda, facilitar uma busca ampla sobre um determinado assunto (LANCASTER, 2002).

Existem sistemas de informação que aplicam a linguagem natural. Tal linguagem nada mais é do que a do discurso comum, por exemplo, utilizada por autores de um determinado campo temático, em que existe um vocabulário ilimitado (LANCASTER, 2002).

A aplicação da linguagem natural na indexação proporciona algumas vantagens, tais como: exatidão ao nomear pessoas e instituições; exaustividade, o que facilita a recuperação da informação; atualização imediata do vocabulário; uso do vocabulário do autor; emprega na busca palavras e frases do autor; além do baixo custo e da facilidade no intercâmbio de material entre base de dados. Por outro lado, verifica-se que a busca requer um esforço

[2] Utilizamos a expressão "linguagem de indexação", de acordo com a corrente teórica inglesa; autores filiados à linha teórica francesa utilizam, como termo correspondente, "linguagem documentária".

intelectual (sinônimos, generalidade, etc.), havendo problemas de sintaxe, com o perigo de associar termos incorretos, e pode acarretar perda de precisão (GIL LEIVA, 2008).

É por isso que, ainda assim, defende-se a aplicação de vocabulário controlado, pois, apesar de possuir algumas falhas, suas vantagens permitem argumentar pela necessidade de vocabulário controlado nos sistemas de informação.

Como desvantagens verificam-se: a lacuna na especificidade; lacunas de exaustividade, pois não é possível contemplar todos os termos, como na linguagem natural, e, além disso, os indexadores podem cometer erros de omissão; a atualização periódica é difícil; pode haver perdas, se as palavras do autor forem distorcidas, e ocorrer atribuição de termos errôneos; para realizar a busca deve-se conhecer a linguagem controlada; alto custo; e dificuldade no intercâmbio de materiais, devido à incompatibilidade entre linguagens controladas (GIL LEIVA, 2008).

Por outro lado, as vantagens são: facilidade na busca (sinônimos, notas de aplicação, hierarquias, associação, etc.); superação de problemas de sintaxe, tais como os problemas envolvendo termos compostos; em níveis normais de indexação, contorna-se a perda completa de precisão; além da vantagem nas bases de dados e nos sistemas multilíngues (GIL LEIVA, 2008).

Lancaster (2002) verifica que o custo e o esforço na aplicação do controle de vocabulário encontram-se na entrada dos dados, ao passo que, nos sistemas de linguagem natural, os esforços se concentram na saída, ou seja, na análise da recuperação de informação.

Lara (2009) verifica, portanto, que a linguagem natural caracteriza-se como uma linguagem dinâmica, cujos lexemas não têm significação unívoca; a língua não tem uma função específica, mas funciona em muitos contextos e para diferentes objetivos, é pouco formalizada e possui alto poder combinatório e, além disso, é utilizada para falar de si mesma.

A linguagem de indexação, por outro lado, possui significados fixados visando à univocidade de interpretação — embora não se possa garanti-la totalmente. É caracterizada também por ser estática, relativamente formalizada, com baixo poder combinatório, e elaborada para desempenhar uma função particular: representar a informação visando à sua recuperação; além disso, é necessário usar a linguagem natural para falar da linguagem de indexação (LARA, 2009).

Embora a linguagem natural e a linguagem de indexação sejam distintas em alguns aspectos, na indexação estabelecem-se relações entre essas linguagens, justamente porque a

linguagem natural é a expressão do documento original e da necessidade informacional do usuário. A linguagem de indexação é utilizada na tradução do conteúdo expresso em um documento em forma de enunciados singulares para possível recuperação e que, advindos da linguagem natural, passam por um processo de normalização prévia para constituir uma sintaxe particular em um campo semântico previamente determinado (MOREIRO GONZÁLEZ, 2004).

Fugmann[3] (1982 *apud* RIVIER,1992), assim como Muddamalle[4] (1998 *apud* LOPES, 2002), considera a linguagem de indexação e a linguagem natural como complementares, verificando a possibilidade de serem usadas em conjunto para melhorar a recuperação da informação. Tanto uma como outra podem ser aplicadas na indexação de acordo com a espécie de conceito de que se trata. Quando, por exemplo, tratam-se de conceitos individuais ligados a um único objeto, expresso na linguagem natural através de uma única expressão léxica, é possível utilizar a linguagem natural, pois existe um vocabulário bem estabelecido na área, o que pressupõe controle terminológico.

Diferentes concepções influenciaram o desenvolvimento de linguagens de indexação, ainda com reflexos na elaboração e na aplicação desses instrumentos atualmente. Por isso, apresentamos o panorama das linguagens de indexação, desde os primórdios, com a elaboração dos cabeçalhos de assunto e das listas de vocabulário livre, até o desenvolvimento de tesauros.

2.1 Dos cabeçalhos de assuntos aos tesauros

A indexação é um processo em que estão envolvidos aspectos decisivos à boa recuperação da informação. Entre esses aspectos, devemos considerar a consistência na indexação. A consistência, entendida como a "[...] extensão com que há concordância quanto aos termos a serem usados para indexar o documento" (LANCASTER, 2004, p. 68), só pode ser obtida se forem utilizados instrumentos capazes de controlar o vocabulário, permitindo que, entre diversas possibilidades, seja utilizado um único termo para a representação da informação.

Gil Leiva (2008) relaciona a qualidade da indexação ao desenvolvimento de um processo de indexação consistente, sendo que um dos instrumentos que se tornam

[3] FUGMANN, Robert. The complementarity of natural and indexing laguages. *International Classification*, v.9, n.3, p.140-144, 1982.
[4] MUDDAMALLE, Manikya Rao. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. *Jasis*, v. 49, n. 10, p. 881-887, 1998.

indispensáveis para garantir essa consistência é a linguagem de indexação, como as listas de descritores, listas de cabeçalhos de assunto e tesauros. Isso porque cada conceito tem uma única representação terminológica que o indexador empregará tanto na indexação da informação como na sua posterior recuperação. “Uma indexação consistente repercute de uma forma positiva no sistema porque evita que documentos que tratam dos mesmos temas apareçam sob indexações diversas” (GIL LEIVA, 2008, p. 115, tradução nossa).

Nesse sentido, verificamos que as linguagens de indexação contribuem para oferecer qualidade ao processo de indexação. A evolução das linguagens de indexação nos revela que nem todos os aspectos foram contemplados desde que surgiram no fim do século XIX. Dessa forma, procuramos explorar, ainda que sucintamente, a evolução e as características de cada tipo de linguagem de indexação.

Apresentamos, inicialmente, os cabeçalhos de assuntos, voltados ao tratamento de assuntos de caráter mais genérico, e os tesauros, de caráter mais especializado, com preocupações voltadas ao estabelecimento de relações semânticas entre conceitos.

2.1.1 Os cabeçalhos de assuntos

Os cabeçalhos de assuntos surgiram nos Estados Unidos, concebidos como instrumentos para catalogação de assuntos de bibliotecas. Como primeiro tipo de linguagem de indexação, foram desenvolvidos apenas para padronizar a entrada dos assuntos de catálogos bibliográficos ou índices (NOVELLINO, 1996). Muitas bibliotecas possuíam seus catálogos, constituídos apenas por fichas catalográficas que ofereciam o acesso às obras por autor e título. A necessidade de pesquisar obras por assuntos específicos caracteriza o início da elaboração de instrumentos de representação temática da informação.

Alguns dos fatores que determinaram o surgimento dos cabeçalhos de assuntos estão relacionados ao fato de que apenas os títulos não representavam adequadamente os assuntos das obras. Havia problemas relacionados às subdivisões de assunto, ou seja, as obras com mais de um assunto e os livros de assuntos relacionados não podiam ser identificados, assim como ocorria com a identificação das obras que relacionavam os assuntos a lugares e épocas diferentes (CESARINO; PINTO, 1978).

Os cabeçalhos de assuntos constituem um tipo de linguagem pré-coordenada, de estrutura associativa ou combinatória que consiste em uma lista alfabética de palavras ou expressões da linguagem natural, que, normalizadas, são capazes de representar os temas de

que trata um documento e por meio dos quais se recuperam os documentos do acervo (GIL URDICIAN, 2004; GIL LEIVA, 2008).

Os princípios que orientaram o desenvolvimento dos cabeçalhos de assuntos foram cunhados por Charles Ammi Cutter, em sua obra "*Rules for a Dictionary Catalog*", publicada em 1876. Entre os princípios mais difundidos estão os de especificidade, o princípio de uso ou de garantia literária, o princípio sindético e o princípio da entrada direta. Vejamos cada um destes princípios:

QUADRO 2 - Princípios de Cutter

PRINCÍPIOS DOS CABEÇALHOS DE ASSUNTOS	
Princípio de especificidade	Sugere que a entrada de um assunto deve ser realizada por um termo mais específico, em detrimento de um mais geral.
Princípio de uso	Indica que os assuntos serão definidos em função de como serão buscados por usuários daquela biblioteca.
Princípio sindético	Indica que deve ser elaborada uma rede bem construída de referências cruzadas para ajudar ou mesmo para superar o problema de aproximações absurdas de assunto e a separação de assuntos relacionados.
Princípio da entrada direta	Indica que, ao utilizar termos compostos, ou seja, termos formados por mais de uma unidade lexical, estes devem ser indicados na forma em que se apresentam na linguagem natural, sem a inversão das palavras.

Fonte: CESARINO e PINTO (1978); MARTINHO (2010)

Esses princípios foram muito importantes, por nortear a construção da representação de assuntos nos catálogos das bibliotecas, que antes era realizada de acordo com o julgamento do próprio catalogador (CESARINO e PINTO, 1978).

Mas, ainda que os cabeçalhos de assuntos tenham sido elaborados com base nesses princípios, existem muitas críticas sobre a sua aplicação, principalmente quanto à sua sistematização (TÔRRES, 2011; GOMES; MARINHO, 2011).

De acordo com Tôrres (2011) os cabeçalhos de assuntos, como instrumentos de representação e recuperação da informação, apresentam uma série de inconsistências em sua construção e em seu uso. Tôrres (2011) tem tratado especialmente da sintaxe dos cabeçalhos de assuntos, que diz respeito à combinação dos elementos ou palavras que formam os cabeçalhos compostos (que são aqueles constituídos por mais de uma unidade lexical).

Tôrres (2011) observou que, embora as regras de Cutter representem um marco importante na história da catalogação de assuntos, os princípios não fornecem diretrizes consistentes para a elaboração da sintaxe dos termos compostos e para a determinação do seu ponto de acesso. São baseadas na linguagem natural, especialmente na sintaxe da língua inglesa, e fundamentadas em uma prática orientada pelo tradicional "bom senso" do profissional catalogador ou indexador (TÔRRES, 2011).

É importante associar determinadas concepções à época em que foram empregadas. Quando foram elaboradas as regras que fundamentaram os cabeçalhos de assuntos, o conhecimento produzido não apresentava o nível de complexidade de hoje. Esse fato é constatado, por exemplo, na pesquisa realizada por Puranik, que analisou a natureza dos assuntos dos artigos científicos. Até 1900, mais de 50% desses assuntos eram simples, e, após 1950, mais de 85% eram compostos. Tal fato talvez explique um dos pressupostos básicos de Cutter, que se encontra subjacente às suas regras: o de que os assuntos geralmente podem ser nomeados por cabeçalhos simples (TÔRRES, 2011).

Outro aspecto que caracteriza os cabeçalhos de assuntos é a ordem alfabética. Segundo Coates, os catálogos alfabéticos de assunto devem ter a dupla função de permitir a identificação de documentos sobre um dado assunto e de possibilitar o acesso aos documentos que tratam de assuntos relacionados. Essas são as duas funções precípuas do catálogo alfabético de assunto. Nesse sentido, a ordem alfabética, que é um mecanismo utilizado na ordenação das entradas desse tipo de catálogo, permite apenas agrupar palavras segundo as letras que as constituem. Portanto, não possibilita o seu agrupamento pelas ideias que designam. A única alternativa nos cabeçalhos de assuntos são as referências ou remissivas (TÔRRES, 2011).

Segundo Cesarino e Pinto (1978), os cabeçalhos de assuntos apresentam as seguintes características:

QUADRO 3 - Características dos cabeçalhos de assuntos

CABEÇALHOS DE ASSUNTOS
a) Linguagens estruturadas e pré-coordenadas que, de certa forma, apresentam limitações na pesquisa
b) Os termos do vocabulário controlado são selecionados de um dicionário existente, o que o caracteriza como sistema fechado
c) Os cabeçalhos de assuntos exercem função prescritiva
d) Linguagens não hierárquicas
e) São enumerativos, oferecendo poucas possibilidades de síntese
f) Arranjo alfabético
g) Linearidade, aplicável apenas a pesquisas unidimensionais.
h) Pouca sistemática na elaboração de cabeçalhos de assuntos e na elaboração de referências cruzadas

Fonte: HORNER (1970)[5] citado por CESARINO e PINTO (1978)

Nesse contexto, verificamos, na FIG. 1, a estrutura de relacionamentos de um cabeçalho de assunto:

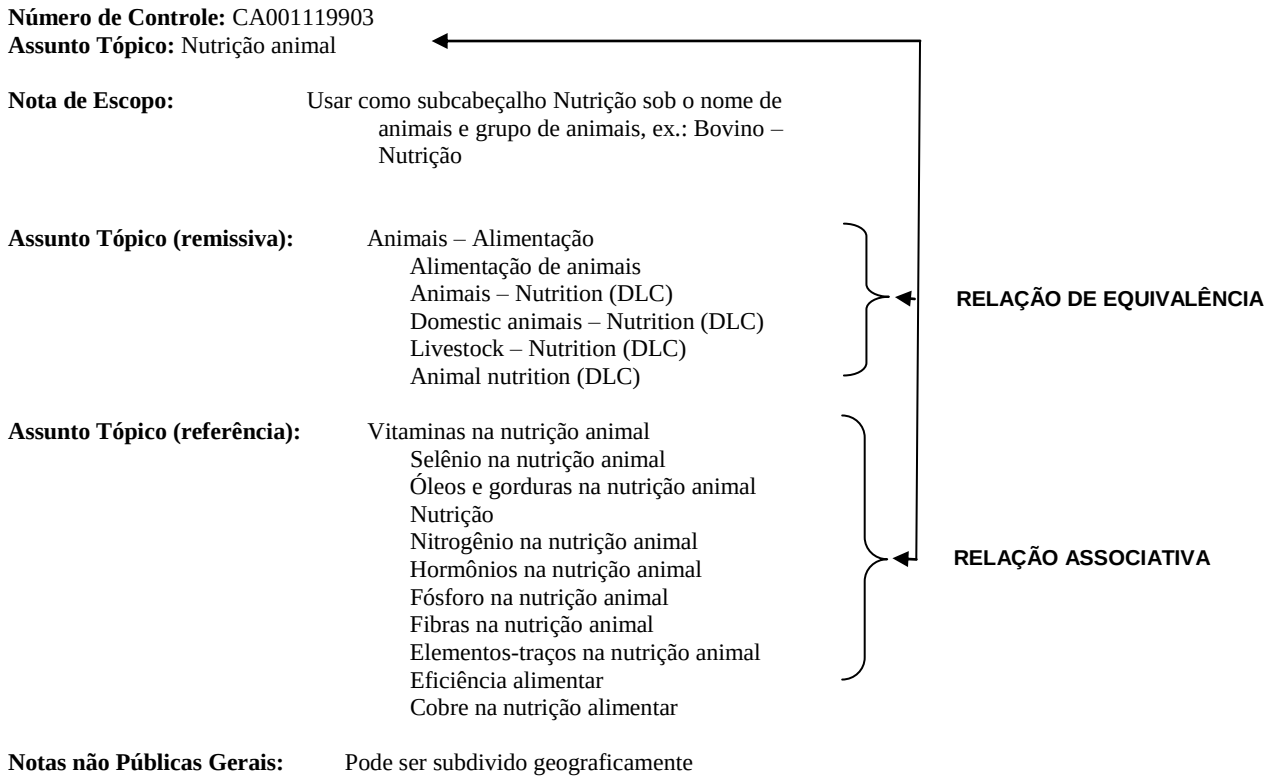


FIGURA 1 - Relacionamentos estabelecidos nos cabeçalhos de assuntos[6]

[5] HORNER, John. Criticism and evaluation of conventional subject headings lists. In: _____. *Cataloguing*. London: Association of Assistant Librarians, 1970. p. 136-42.

Os cabeçalhos de assuntos originaram-se dos assuntos dos catálogos alfabéticos de bibliotecas — caracterizados pela garantia literária —, refletindo, portanto assuntos de acervos particulares. Nesse sentido, refletem assuntos específicos de interesse de comunidades particulares, o que, por vezes, não deve ser generalizado a outras bibliotecas.

Nesse contexto, os critérios estabelecidos ao longo do tempo para a atualização desses cabeçalhos de assuntos são aspectos relevantes, que devem ser considerados até mesmo quando analisamos as atuais linguagens de indexação — que, provavelmente, tiveram influência e ainda apresentam resquícios das concepções vigentes naquela época.

Os cabeçalhos de assunto da *Library of Congress Subject Headings* (LCSH) por exemplo, são produto da acumulação de cabeçalhos incorporados desde 1898 (GIL URDICIAIN, 2004). Seu crescimento se deu em função das demandas dos usuários e, portanto, as fases de criação de suas listas tiveram influência de diferentes mentalidades e filosofias. Dessa forma, uma das grandes preocupações verificadas se refere à possibilidade do seu uso por outros sistemas, sem uma análise de adaptação à sua própria realidade (CESARINO; PINTO, 1978; DODEBEI, 2002).

Muitos sistemas adotam o uso de cabeçalhos elaborados em outras instituições. Isso tem implicações profundas sobre a representação da informação, na medida em que existem diferenças culturais, sociais e políticas subjacentes aos instrumentos de representação da informação.

As mudanças ocorridas sob influência dos avanços tecnológicos e científicos, nas ciências em geral e também no campo da documentação, durante o início do século XX e, principalmente, após a Segunda Guerra Mundial, configuraram uma dinâmica na produção do conhecimento. As relações entre temáticas estabelecidas nos sistemas pré-coordenados e hierárquicos se tornaram demasiado rígidas, estáticas, para possibilitar a inclusão de novos conceitos que surgiam a cada dia (CAMPOS, 2001; CURRÁS, 2005).

O desenvolvimento de gêneros documentais, como artigos de periódicos e relatórios de pesquisa, se intensifica, sendo constituídos por conteúdos bastantes especializados. Os instrumentos de representação de caráter mais geral, os "sistemas de classificação" e os "cabeçalhos de assuntos", não possuíam as condições necessárias para representar os conteúdos desses textos técnico-científicos (DODEBEI, 2002).

---

[6] www.fgv.br/bibliodata

A possibilidade de combinar novos termos na busca de informação não era possível dentro de expressões pré-coordenadas, determinadas pelos sistemas vigentes até aquele momento (VIZCAYA ALONSO, 1997). Exigia-se outra solução para atender às novas necessidades de pesquisa surgidas no universo dos usuários.

De fato, nesse período há o início de uma reflexão sobre o documento enquanto suporte e sobre seu conteúdo temático. Foi-se difundindo a importância do conteúdo dos documentos, independentemente do suporte, que se convertia em informação útil e necessária em função do seu próprio significado e da demanda para posterior uso (CURRÁS, 2010).

Nesse sentido, foi proposto um instrumento que contemplasse as novas exigências para organização, representação e recuperação da informação, que vem a ser conhecido como "tesauro".

2.1.2 Os tesauros

Expresso em latim como "*thesaurus*", o termo tesauro é de origem grega ("*thesaurós*") e significa "tesouro", no sentido de "armazenagem" ou "repositório de palavras" (VICKERY, 1960[7] *apud* DODEBEI, 2002). Ou, como Gil Urdiciain (2004) explica, "tesaurizar" quer dizer, "acumular riquezas" e, no sentido figurativo, significa "acumular bens intelectuais".

Seu uso para designar um tipo de linguagem de indexação sucede da publicação da obra "*Thesaurus of English Words and Phrases*" por Peter Mark Roget, em 1852. Nesse dicionário, parte-se do significado, uma ideia, para chegar a todas as palavras que o representam, ao contrário dos tradicionais dicionários, nos quais se parte de uma palavra para encontrar seu significado (RIVIER, 1992; CAMPOS, 2001; DODEBEY, 2002; GIL LEIVA, 2008).

O Thesaurus de Roget é composto de duas partes: na primeira, existe uma estrutura classificatória de ideias constituída por diversas categorias que são subdividas em tópicos. A segunda parte é constituída por um índice alfabético, que apresenta a associação entre os cabeçalhos (sob os quais ocorrem as palavras e frases) e os números, que representam as ideias na parte sistemática (CAMPOS, 2001).

Nesse período de transição dos cabeçalhos de assuntos para os tesauros, surge, em 1951, nos Estados Unidos, o sistema Unitermo, introduzido por Mortimer Taube. O sistema

[7] VICKERY, B. C. Thesaurus: a new world in documentation. *Journal of Documentation*, [S.l.], v. 16, n. 4, p. 181-89, dez. 1960.

Unitermo, segundo Lancaster (1986, p. 31), tem como principal característica "a representação do assunto por palavras únicas extraídas do texto de um documento sem nenhuma forma de controle".

Em 1951, os computadores começavam a ser utilizados e acreditava-se que a aplicação de unidades isoladas não oferecia inconveniente, visto que, no processo de busca, era possível combinar essas unidades utilizando o sistema Booleano. Desse modo, é nesse período também que as expressões "recuperação da informação" ("*Information Retrieval*") e "palavra-chave" (que dá origem a "descritor"), também cunhadas por Taube, se popularizam (CURRÁS, 2005).

Considerando que as palavras do sistema Unitermo não tinham uma forma de controle, posteriormente exigiu-se a formalização de termos autorizados (descritores). Ou seja, sentiu-se a necessidade de controle de vocabulário e de uma estruturação segundo diferentes relações semânticas que permitissem, no momento da busca, alcançar expressões linguísticas de maior profundidade semântica para não somente uma recuperação de informação de maior nível de precisão, mas, também, de maiores níveis de relações intra e interdisciplinares (VIZCAYA ALONSO, 1997).

De fato, a aplicação de um único termo e a ausência de controle de vocabulário ocasiona problemas à consistência na indexação.

Existem fatores intervenientes que nos fazem refletir sobre a necessidade de estabelecer controle de vocabulário nos sistemas de representação e de recuperação de informação. Segundo Cesarino e Pinto (1978), entre esses fatores verificam-se os fatores humanos relacionados às diferenças de cultura, de experiência dos autores e de domínio da terminologia entre indexadores, autores e usuários. Além disso, existem fatores referentes à própria linguagem natural, suscetível aos fenômenos linguísticos de sinonímia, polissemia e sintaxe, e fatores hierárquicos, em que um conceito implica em outros mais amplos e/ou mais restritos. Por isso, é necessário prudência quanto à contextualização dos conceitos na representação dos documentos para permitir acesso preciso à informação.

Os primeiros tesauros foram concebidos com ordenação alfabética, mas as deficiências dessa forma de arranjo evidenciaram a necessidade de incluir uma abordagem sistemática para estabelecer relações entre conceitos.

Segundo Foskett[8] (1985 *apud* CAMPOS, 2001), em 1950 Luhn já utilizava o termo "*Thesaurus*" para nomear seu sistema de palavras que possuía uma estrutura de referências cruzadas. Ao invés de utilizar uma estrutura de listagem alfabética, Luhn percebeu que era necessário evidenciar as noções que ligavam uma palavra a outras, estabelecendo relações entre elas.

As relações semânticas estabelecidas em um tesauro permitem expressar relações de equivalência, hierárquicas e de associação que, estruturadas de forma alfabética e/ou sistemática e/ou gráfica e, sinalizadas por código especiais (USE, UP, TG, TE, TR), lhe garantem uma organização própria que facilita o processo de uso para representação e recuperação da informação.

Vejamos um exemplo que ilustra as relações semânticas contempladas por um tesauro:

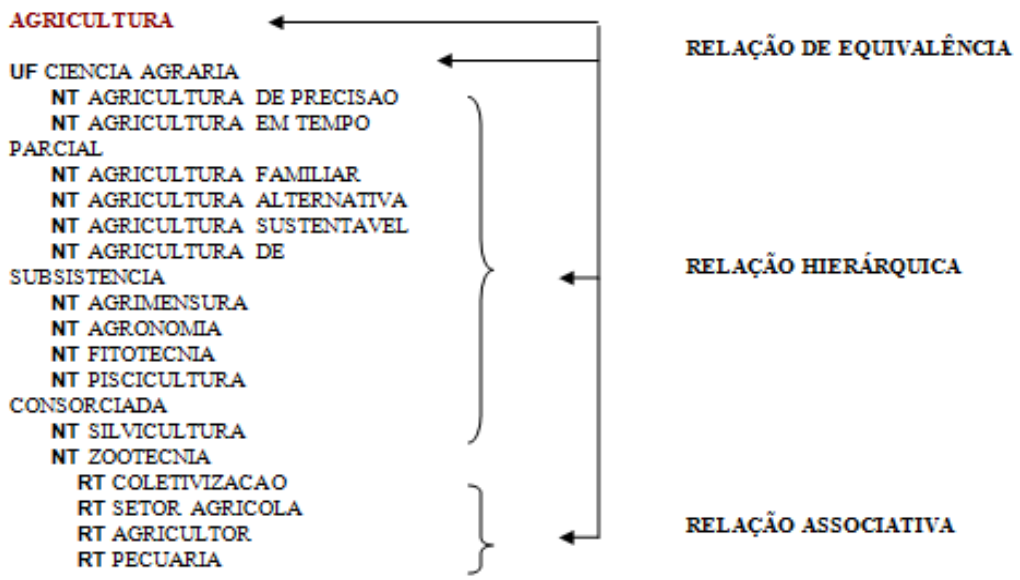


FIGURA 2 - Relacionamentos entre termos estabelecidos nos tesauros
Fonte: Adaptado do THESAGRO (Thesaurus Agrícola Nacional)[9]

Na FIG. 2, o descritor "AGRICULTURA" estabelece relação de equivalência com o não descritor "CIENCIA AGRARIA", indicado pelo código UF ("*used for*" ou "usado para"). É possível visualizar, também, a relação de hierarquia entre o descritor genérico "AGRICULTURA" e os descritores específicos, identificados pelo código NT ("*narrower term*" ou "termo específico"), assim como a relação associativa entre o descritor

[8] FOSKETT, D. Thesaurus. In: *Subject and information analysis*. New York: M. Dekker, 1985.
[9] Disponível em: <http://www.agricultura.gov.br>

"AGRICULTURA" e os descritores que apresentam o código RT ("*related term*" ou "termo relacionado").

As relações de equivalência são estabelecidas entre termos sinônimos para evitar a incompatibilidade entre a linguagem do sistema e a do usuário. Possibilitam, também, considerar a coincidência de significado entre um termo antigo e um termo novo; um termo popular e seu correspondente cientifico; um termo geral e um termo específico utilizado em uma região; e entre termos de diferentes origens etimológicas (MOREIRO GONZÁLEZ, 2004). Possibilitam considerar, ainda, as relações entre quase sinônimos e entre termos intimamente relacionados e que, para os propósitos do tesauro, são considerados sinônimos — tais como, por exemplo, termos que representam diferentes pontos de vista da mesma propriedade (ESTABILIDADE/INSTABILIDADE, NUTRICAO/DESNUTRICAO), ou termos que têm superposição significativa (GENÉTICA/HEREDITARIEDADE), sendo necessário definir um termo preferido e remeter aos outros (AITCHISON, GILCHRIST, 1979).

Para Cintra *et al*. (2002), as relações de equivalências são importantes porque podem controlar as variações de significado, permitindo maior rigor no tratamento da informação e eficácia na recuperação da informação.

As relações hierárquicas, segundo Currás (2005), permitem reunir os descritores estabelecendo relacionamentos entre termos superiores-genéricos e termos subordinados-específicos. Sendo assim, são estabelecidas as relações genéricas indicando que todo conceito pertencente à categoria do conceito específico (a espécie) faz parte da extensão do conceito amplo (o gênero). Logo, um conceito específico possui todas as características do conceito mais amplo e, pelo menos, uma característica distintiva adicional que serve para diferenciar conceitos específicos no mesmo nível de abstração (CINTRA *et al*., 2002).

Por sua vez, as relações associativas se estabelecem quando as famílias, ou grupos de termos afins, são estudadas no plano horizontal, considerando diferentes pontos de vista (CURRÁS, 2005).

Desse modo, constata-se o valor que adquire a adequada determinação de relacionamentos na estruturação dos tesauros, uma vez que essa composição reflete a organização de um domínio, pelo qual são construídos novos conhecimentos e relações semânticas.

Cabe salientar que, ao comentar sobre relações semânticas, atualmente se verifica outra questão importante que se refere à interoperabilidade entre vocabulários controlados. De

acordo com Soler Monreal (2009), a interoperabilidade consiste em desenvolver métodos que permitam utilizar vocabulários controlados em múltiplas bases de dados e sistemas, permitindo compartilhá-los por indexadores e buscadores, incluídos, também, os vocabulários controlados multilíngues.

Dessa forma, constata-se que as relações semânticas se dão entre os conceitos de um determinado tesauro, mas também entre tesauros, assim como pode haver relações entre áreas diferentes, uma vez que áreas do conhecimento se desenvolvem a partir do esforço de investigação interdisciplinar. A interoperabilidade envolvendo vocabulários multilíngues é fundamental para que haja um diálogo entre distintas comunidades linguísticas.

De acordo com a norma da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) (1973, p. 6) "tesauro" é "um vocabulário controlado e dinâmico de termos relacionados semântica e genericamente cobrindo um domínio específico do conhecimento", que serve como um "dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem do sistema (linguagem de documentação, linguagem de informação) mais restrita".

Dito de outro modo, o tesauro pode ser entendido como uma linguagem constituída de um vocabulário controlado formado por descritores (termos autorizados para indexação) e não descritores, em que são estabelecidas relações semânticas que permitem, por um lado, descrever o conteúdo temático de um documento e, por outro, construir as expressões de busca para recuperação da informação.

Para Gil Urdiciain (2004), os tesauros são linguagens com uma série de vantagens, destacando-se a sua flexibilidade e a sua capacidade de especialização, que permitem estabelecer entre os termos de seu vocabulário uma multiplicidade de combinações, bem como o alto nível de controle terminológico e a facilidade de revisão.

Sob o ponto de vista das correntes teóricas que originaram os tesauros, verificamos, de um lado, os tesauros elaborados na América do Norte, de abordagem alfabética, decorrência do desenvolvimento de cabeçalhos de assunto para o sistema Unitermo, e, de outro, uma vertente europeia, de abordagem sistemática concentrada, sobretudo, nas investigações do *Classification Research Group* (CRG), sob forte influência da Teoria da Classificação, de Ranganathan (CAMPOS, 2001).

No entanto, com relação ao estabelecimento de bases teóricas para a determinação das unidades que constituem os tesauros, as duas vertentes ainda não haviam resolvido essa

questão, que vem a ser tratada apenas na década de 1970, com a Teoria do Conceito, de Dahlberg.

A Teoria do Conceito estabelece um método para a fixação dos conceitos e para o seu posicionamento em um sistema conceitual (CAMPOS, 2001). Tanto na vertente norte-americana quanto na europeia (acima citadas), a palavra ou o termo é considerado como unidade que constitui o tesauro, ao passo que, na Teoria do Conceito, considera-se que essa unidade é o conceito. Compreende-se que o conceito é constituído pelo conjunto de atributos que caracterizam um objeto e, nesse caso, o termo designa o conceito, sendo que o que permite defini-lo em um sistema de conceitos são esses seus elementos, ou seja, as suas características.

A Teoria do Conceito ofereceu uma base teórico-metodológica importante para sustentar a construção de tesauros, visto que, até aquele momento, poucos avanços com relação à definição de conceitos eram encontráveis na literatura de Ciência da Informação. Até aquele momento, a unidade de trabalho continuava sendo a palavra.

Para desenvolver a Teoria do Conceito, Ingetraut Dahlberg fundamentou-se nas contribuições da Terminologia, oferecendo aporte teórico para a pesquisa e o desenvolvimento de tesauros e originando a tendência de pesquisa que se conhece por “tesauros terminológicos” ou “conceituais”.

Os tesauros conceituais são compreendidos como tesauros com base em conceitos, em que são instituídos princípios para o estabelecimento do termo/conceito e das relações entre eles. A Teoria do Conceito e a Teoria da Classificação Facetada de Ranganathan contribuem para a elaboração de tesauros conceituais, estabelecendo bases para a identificação dos conceitos, dos termos e das relações entre eles, e, ainda, para a sua ordenação sistemática (CAMPOS; GOMES, 2006).

É possível verificar a dificuldade de estabelecer os limites entre, de um lado, o controle do vocabulário, a flexibilidade inerente à linguagem, a capacidade de expressão relacionada à representação da informação e a interface desses aspectos, com, de outro, principalmente a ação de comunicar.

Nesse sentido, apresentamos, a seguir, características das duas vertentes de desenvolvimento dos tesauros e características do surgimento dos tesauros conceituais.

QUADRO 4 - Características das vertentes sobre tesauros

TESAUROS			
	Vertente norte-americana	**Vertente europeia**	**Tesauro conceitual**
	Unitermo	Thesaurofacet	Tesauro baseado em conceito
Características	Ruptura aos cabeçalhos de assuntos	Influência da Teoria da Classificação de Ranganathan	Terminologia e Teoria do Conceito de Dahlberg
	Uso de palavra única	Categorização	Categorização
	Ausência de controle de vocabulário	Controle de vocabulário	Definição do conceito
	Abordagem alfabética	Abordagem sistemática	Controle de vocabulário
	Sistema pós-coordenado	Sistema pós-coordenado	Abordagem sistemática
	Evolução pragmática	Estabelecimento de relações semânticas	Sistema pós-coordenado
	Levantamento do domínio	Levantamento do domínio	Estabelecimento de relações semânticas
			Levantamento do domínio

Fonte: Elaborado pela autora de acordo com CAMPOS (2001); CAMPOS *et al*. (2006).

Dentre as características analisadas, o que distingue a vertente norte-americana da vertente europeia é o uso de palavras simples, a ausência de controle de vocabulário e a estruturação segundo uma ordem alfabética, ao passo que, na vertente europeia, verificou-se a necessidade de controle de vocabulário e de organização sistemática para criar condições de se estabelecerem relações semânticas de equivalência, hierárquica e associativa, proporcionando a possibilidade de manipular uma estrutura mais flexível. Podemos verificar esse panorama no esquema evolutivo dos tesauros apresentado em seguida:

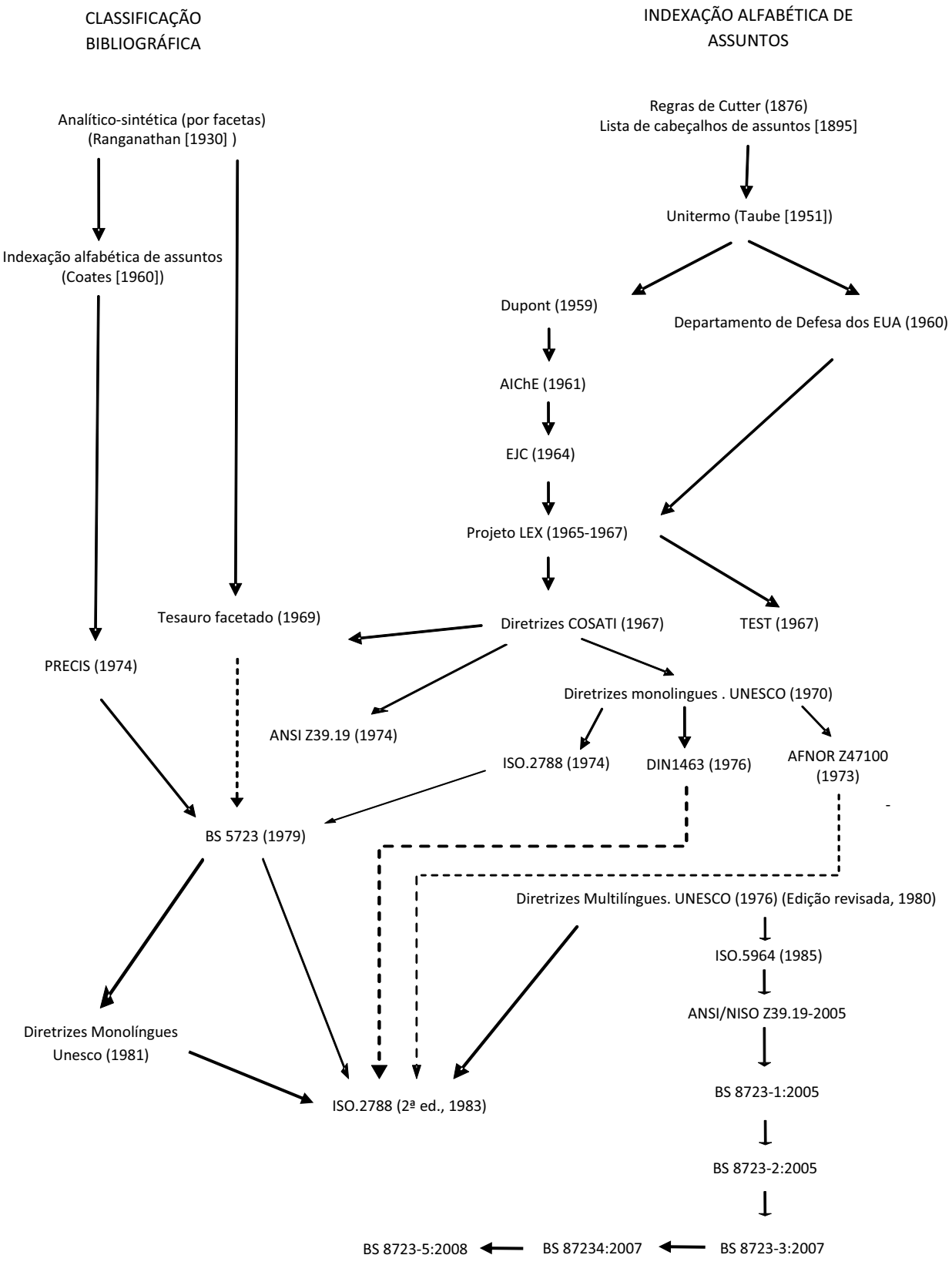


FIGURA 3 - Evolução das normas sobre tesauro (as linhas tracejadas indicam pouca influência)
Fonte: LANCASTER (2002, p.50, tradução e adaptação nossa)

No esquema acima, Lancaster (2002) evidencia as duas linhas que influenciaram o desenvolvimento dos tesauros. Apresenta as influências da vertente norte-americana desde os princípios de Cutter para a elaboração dos cabeçalhos de assuntos para aplicação em catálogos de assuntos de bibliotecas até a necessária ruptura no pós-guerra frente à especialização do conhecimento e ao desenvolvimento de computadores, surgindo o sistema Unitermo, proposto por Mortimer Taube. Este sistema, caracterizado pelo uso de palavras simples, manifestou essa característica também nos primeiros tesauros elaborados nos Estados Unidos e influenciou a elaboração das primeiras normas de construção de tesauros (LANCASTER, 2002). Nos Estados Unidos, a iniciativa de desenvolvimento de tesauros esteve sob responsabilidade, principalmente, de órgãos ligados ao governo, em áreas especializadas como engenharia, etc.

Por outro lado, uma linha de pensamento se constituiu na Europa, no Reino Unido, buscando alicerce teórico para o desenvolvimento dos tesauros sob influência da Teoria da Classificação Facetada de Ranganathan. E, posteriormente, nas décadas de 1970 e 1980, com o aparecimento da Teoria do Conceito e de outras teorias terminológicas, as discussões em torno de metodologias, diretrizes e teorias para a fundamentação do desenvolvimento de tesauros são amplamente pesquisadas. Verificamos, portanto, a profusão de tesauros em várias áreas do conhecimento (engenharia, química, área agrícola, medicina).

Reconhecemos, portanto, que os princípios subjacentes a essas vertentes influenciaram também a elaboração dos atuais vocabulários controlados, na medida em que se buscaram os melhores referenciais de cada vertente para a elaboração das normas internacionais (ISO 2788-1986) de elaboração de tesauros na década de 1980 e as posteriores.

Desse modo, procuramos sintetizar os traços dos cabeçalhos de assuntos e dos tesauros, no intuito de evidenciar a evolução que ocorreu desde o final do século XIX:

QUADRO 5 - Características dos cabeçalhos de assuntos e tesauros

Cabeçalhos de assuntos	**Tesauros**
Linguagem pré-coordenada	Linguagem pós-coordenada
Organização alfabética	Organização sistemática
Influência dos princípios de Cutter: da especificidade, do uso, sindético e da entrada direta	Influências da Teoria da Classificação e da Teoria do Conceito de Dahlberg

(continua)

(conclusão)

Cabeçalhos de assuntos	Tesauros
Relações apenas por referências cruzadas do tipo “ver” e “ver também”	Relações semânticas associativas, hierárquicas e de equivalência
Entrada direta	Apresentam símbolos consensuais para descrever as relações semânticas
Assuntos gerais de catálogos de bibliotecas	Assuntos de áreas especializadas
Para catálogos de assuntos	Para sistemas de indexação e recuperação da informação
Estrutura rígida	Estrutura flexível
Terminologia de uso da linguagem natural do usuário	Alto nível de controle terminológico

Fonte: Elaborado pela autora

Verificamos que os cabeçalhos de assuntos foram alvo de várias críticas. Embora os princípios de Cutter orientassem a elaboração de cabeçalho de assuntos, também possuíam vários aspectos sem esclarecimentos. No contexto de aplicação em catálogos de assuntos de bibliotecas, a característica de pré-coordenação se tornava uma alternativa para que não fosse necessário gerar uma quantidade grande de fichas catalográficas. No entanto, se analisarmos as exigências a partir da década de 1950, não era possível continuar utilizando cabeçalhos de assuntos em sistemas automáticos de recuperação de informação.

Foi nesse sentido que o sistema Unitermo tornou-se uma alternativa ao modelo anterior, caracterizado pela coordenação dos assuntos. O sistema Unitermo ofereceu a alternativa de utilizar termos constituídos apenas por uma palavra e sem um controle rígido do vocabulário.

Porém, se por um lado o sistema Unitermo ofereceu a flexibilidade para atribuir qualquer palavra na indexação, por outro ocasionou situações problemáticas. Por ser totalmente pós-coordenado, o sistema Unitermo não permitiu o uso de conectivos da língua e são os conectivos que oferecem coerência ao discurso. O sistema sofreu críticas por essa total pós-coordenação acabar gerando uma ampla possibilidade de interpretações que interfere na recuperação da informação. Uma das situações mais problemáticas é causada pela polissemia, fenômeno em que uma palavra pode ter vários significados, como no exemplo da palavra “banco” que pode se referir à entidade financeira, base de dados ou assento.

Com relação à organização do vocabulário, o sistema Unitermo permaneceu com a ordenação alfabética, assim como os cabeçalhos de assuntos, o que era uma limitação tanto para a indexação como, principalmente, para a recuperação da informação. A rede de relações entre conceitos confere ao tesauro uma organização sistemática em que o usuário visualiza e especifica melhor as suas pesquisas, ao mesmo tempo em que permite ao indexador uma possibilidade maior de exploração do vocabulário para a tradução dos conceitos para representação.

A diferença entre um tesauro e os cabeçalhos de assuntos, é que, nos cabeçalhos de assuntos, os termos que os compõem são relacionados a priori, em um processo de pré-coordenação que lhes confere certa rigidez. Em um tesauro, os termos, simples ou compostos, estão relacionados entre si de forma que permitam combinações em um processo de pós-coordenação. São, portanto, mais flexíveis e sua atualização se torna mais dinâmica e rápida (CURRAS, 2005).

Além disso, verificamos que a influência da Teoria da Classificação e da Teoria do Conceito na elaboração de tesauros fundamentou a exploração do conceito como unidade de representação, o que permitiu a organização em um sistema de conceitos — ao contrário dos vocabulários anteriores, em que a unidade de representação é a palavra e a organização é alfabética.

2.2 Ontologias

Os meios de armazenamento, organização e acesso às informações sofreram grandes mudanças ao longo de toda a história humana e, durante a segunda metade do século XX, não foi diferente. Houve grandes avanços na aplicação de Tecnologias de Informação e Comunicação (TICs), propiciadas também pelo desenvolvimento da Web no início da década de 1990 por Tim Berners-Lee.

A possibilidade de criar, modificar e buscar conteúdos transformou a Web em um ambiente, de certa forma, caótico e que, ao mesmo tempo, pode potencializar a construção de um conhecimento coletivo.

Entretanto, é necessário que haja esforços para criar uma estrutura capaz de organizar e dispor todo esse conhecimento produzido para que, de fato, as informações disponíveis na web sejam recuperadas e tornem-se úteis às pessoas.

Nesse sentido, surge em 2001 o projeto "Web semântica" do *World Wide Web Consortium* (W3C) em que considera a Web semântica como uma extensão da Web atual e não uma web separada, na qual é dado à informação um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação, ou seja, que a informação possa ser compreendida tanto por humanos como por máquinas (BERNERS-LEE, T., LASSILA, O., HENDLER, J., 2001).

Dessa forma, uma nova categoria de instrumento de representação do conhecimento, conhecida como "ontologia", começa a despertar interesse na Ciência da Informação no final da década de 1990.

As ontologias formais escritas em linguagem de máquina surgem no campo da Inteligência Artificial, no início de 1990, como recurso de representação do conhecimento, mas o conceito tem origem no campo da Filosofia.

A palavra ontologia tem origem grega, em que "*ontos*" significa "ser" e "*logos*" significa "palavra". Entendida na Filosofia como o estudo ou ciência do "Ser" enquanto "Ser", a ontologia é o estudo da existência de todos os tipos de entidades, abstratas ou concretas, que constituem o mundo (LIMA-MARQUES, 2006). Em alguns tratados de filosofia, considera-se a Ontologia como o estudo do que existe e do que admitimos que existe, para conseguir uma descrição coerente da realidade (CURRÁS, 2010).

Na área de Inteligência Artificial, segundo Lima-Marques (2006), uma das mais fortes razões para o desenvolvimento de ontologias é a possibilidade de compartilhamento e reutilização de conhecimento formalmente representado para uso em sistemas computacionais, o que exige a definição de um vocabulário comum para a representação do conhecimento.

Em uma das definições mais presentes na literatura, compreende-se que "ontologia" é "uma especificação formal e explícita de uma conceitualização compartilhada" (BORST, 1997, p. 12), em que "formal" significa legível para computadores; "especificação explícita" está relacionada a conceitos, propriedades, axiomas explicitamente definidos; "compartilhado" seria o conhecimento consensual; e "conceitualização" diz respeito a um modelo abstrato de algum fenômeno do mundo real (ALMEIDA, 2003).

Ramalho (2010) explica que, na área da Ciência da Informação, uma ontologia pode ser definida como um sistema de representação do conhecimento que possibilita descrever formalmente as propriedades e relacionamentos de um determinado modelo conceitual,

favorecendo a realização de inferências automáticas nos processos de organização e recuperação de recursos informacionais.

Entre os aspectos que se destacam nas ontologias, um dos principais é a sua capacidade de permitir a representação de uma visão de mundo, potencializando as relações semânticas que não poderiam ser obtidas por descrições textuais (RAMALHO, 2010).

Segundo Ramalho (2010), as ontologias possibilitam ir além da representação dos aspectos descritivos e temáticos dos documentos, fornecendo subsídios computacionais para a representação dos próprios domínios, contribuindo para a contextualização das informações.

As ontologias são constituídas por "*classes*" e "*subclasses*" que agrupam um conjunto de elementos de acordo com suas similaridades; por "*propriedades descritivas*", em que as características das classes são descritas; por "*propriedades relacionais*", que tratam dos relacionamentos entre classes de uma mesma hierarquia ou não, descrevendo os tipos de relações existentes; por "*regras e axiomas*", que são enunciados lógicos que impõem condições, possibilitando inferências automáticas; por "*instâncias*", que indicam os valores das classes e subclasses; e por "*valores*", que atribuem valores concretos às propriedades descritivas, indicando os formatos e os tipos de valores aceitos em cada classe (RAMALHO, 2010).

Para desenvolver um ambiente web bem estruturado criou-se a necessidade de que as pesquisas retornem aos princípios básicos para a construção de instrumentos de representação, tais como sistemas de classificação, vocabulários controlados e tesauros. Os aportes teóricos que fundamentam a elaboração de instrumentos tradicionalmente desenvolvidos na Ciência da Informação podem oferecer subsídios teóricos e metodológicos para a construção de ontologias.

Nesse sentido, verifica-se que a estrutura e a concepção das ontologias se distinguem dos tradicionais instrumentos de representação — as linguagens de indexação — apresentados no início deste capítulo, apesar de possuírem aspectos comuns que as aproximam[10]. Podemos verificar no QUADRO 6 algumas das características em relação aos tesauros:

[10] Os aspectos que aproximam e distinguem os tesauros das ontologias são apresentados em Currás (2005); Sales e Café (2008); Gil Leiva (2008); Ramalho (2010); Soler Monreal; Gil Leiva (2010).

QUADRO 6 - Semelhanças e diferenças entre tesauros e ontologias

	Tesauros	Ontologias
Objetivo	Representar e buscar informação	Organizar, explorar, compartilhar e reutilizar informação
Origem	Década de 1950	Década de 1980
Cobertura	Restrita a um campo do saber	Restrita a um âmbito do saber ou setor (econômico, sanitário, educativo, de mercado de trabalho, etc.)
Entorno	Analógico e digital	Digital
Fontes	Autorizadas (literatura científica e linguagem dos usuários)	Autorizadas (literatura cientifica, dados procedimentais, organogramas, causas-efeitos, sintomas-tratamentos, dados estatísticos, etc.)
Linguagem	Linguagem natural e linguagem controlada (terminologia consensual e normalizada)	Linguagem natural, linguagem controlada e linguagem formal
Estrutura	Sistemática ou macrotesauro, hierárquica, alfabética, índice (Kwic ou Kwoc)	Taxonomia, tabela com descritores, relações, atributos, valores, axiomas
Uso de taxonomias **Custo de elaboração**	Não Elevado	Sim Muito elevado
Tipos de relações	Hierárquicas, associativas e de equivalência	Hierárquicas, associativas, de equivalência e qualquer outro tipo (temporais, familiares, causas-efeitos, sintomas-tratamento, etc.)
Inferências	Não	Sim
Definições	Contêm principalmente notas de como empregar um descritor na atividade de indexação e recuperação	Contêm definições universais e consensuais de cada um dos conceitos incluídos na ontologia
Axiomas	Não	Sim. Os axiomas permitem realizar inferências
Reutilização	Toda ou parte da terminologia de um tesauro pode ser integrada em outro mais geral (por exemplo, um tesauro de urbanismo em outro de administração pública); ou também em outro mais específico (por exemplo, um tesauro sobre patrimônio histórico em outro de arqueologia). Custo elevado	As supraontologias (especificações formais do universo) podem ser reutilizadas nas ontologias de âmbito (especificações formais de um âmbito concreto); por exemplo, uma ontologia de medicina pode utilizar parte de uma linguística, como WordNet; ou, também, parte de uma ontologia de economia em outra de comércio eletrônico. Custo médio
Normas	ISO 25964 ANSI/NISO Z39.19 BS8723	Não há um padrão oficial

(continua)

(conclusão)

	Tesauros	Ontologias
Apresentação	Símbolos BT, NT, RT, UF, USE	Recomendações da W3C
Editores de construção	MultiTes, Stride, TCS, Léxico, TermTree 2000, iSGAT, BEAT	Protegé, Ontolingua server, Swoop, OntoEdit
Linguagens de construção	Linguagens de marcação: SKOS-Core, Zthes	Linguagens tradicionais: KIF, Ontolingua, OCML. Linguagens de marcação: OIL, DAM+OILM, RDF, OWL

Fonte: SOLER MONREAL e GIL LEIVA (2010, p. 374, tradução nossa)

Assim como os vocabulários controlados, as ontologias podem ser aplicadas aos sistemas de indexação automática como estruturas de suporte para organização, representação e recuperação da informação, favorecendo, principalmente, a contextualização de informações.

Verificamos, portanto, que, sob uma perspectiva histórica, instrumentos de representação da informação foram elaborados em conformidade com as necessidades e as possibilidades tecnológicas da época e, longe de tornarem-se obsoletas, suas bases teóricas e metodológicas serviram de base para adaptação e aperfeiçoamento dos posteriores instrumentos de representação da informação.

Portanto, é necessário desenvolver reflexões sobre a aplicação de instrumentos de representação da informação tradicionalmente empregados na indexação realizada por humanos e sobre a sua adaptação na indexação realizada por sistemas automáticos. Sendo assim, apresentamos, no próximo capítulo, uma análise dos aspectos que envolvem a indexação automática para que, em seguida, seja possível analisar a aplicação de vocabulários controlados nesse contexto.

3 INDEXAÇÃO AUTOMÁTICA

O conhecimento da sociedade pode ser difundido e servir ao seu desenvolvimento na medida em que existam recursos que viabilizem o registro, a preservação e a disseminação de informações, oferecendo o potencial para gerar novos conhecimentos.

A grande quantidade de informações disponíveis e o favorável contexto que permite a disseminação e a criação de informações tornam complicada a tarefa de buscar com precisão aquilo que se deseja. Por isso, é necessário desenvolver atividades de tratamento da informação para viabilizar a sua representação, através da qual será possível recuperar os documentos.

Nesse sentido, a indexação é um processo fundamental, já que realiza o tratamento do conteúdo temático dos documentos, ou seja, permite criar pontos de acesso por assuntos.

No entanto, realizar a indexação não é uma tarefa fácil; pelo contrário, é um processo complexo, por envolver diversas variáveis relacionadas ao indexador, ao usuário, aos instrumentos de representação de informação, ao documento, ao contexto institucional, etc.

A indexação é compreendida como um processo em que o documento é analisado sob o aspecto de seu conteúdo temático com o objetivo de capturar os conceitos que o representam. Nesse processo de análise considera-se o conteúdo, assim como a importância desta para a comunidade usuária. E, em uma etapa de tradução, os conceitos são representados por termos de uma linguagem de indexação com o objetivo de tornar-se o meio pelo qual os documentos serão recuperados por usuários nos sistemas de informação.

É possível verificar nas definições do conceito de indexação, como apresentado pelos "Princípios de Indexação" do *World Information System for Science and Technology* (UNISIST), de 1981, que o processo se constitui basicamente de dois estágios. No primeiro estágio se estabelecem os assuntos tratados no documento e, no estágio de tradução, os conceitos são expressos em termos de uma linguagem de indexação.

É um processo em que se consideram tanto os objetos suscetíveis de ser representados por conceitos quanto as perguntas dos usuários para, em última instância, satisfazer necessidades de informação (GIL LEIVA, 2008). O conteúdo do documento, assim como a análise das necessidades dos usuários, são fontes de referência ao realizar o processo de

indexação com o objetivo de permitir o armazenamento da informação para atender necessidades informacionais.

Portanto, a indexação se reveste de importância em qualquer sistema de informação, visto ser etapa estratégica em que a qualidade no seu processo implica diretamente nos resultados de recuperação da informação.

A qualidade na indexação está intimamente associada ao estabelecimento de uma política de indexação que considere características de consistência (ZUNDE; DEXTER[11], 1969), exaustividade e especificidade (ROBREDO, 2005) e a ausência de erros associada à correção na indexação (GIL LEIVA, 1999).

Nesse contexto, a política de indexação constitui-se na formalização dos processos, procedimentos, instrumentos e de toda filosofia profissional subentendida nas atividades de indexação que servem como diretriz ao desenvolvimento dessas atividades.

Segundo Carneiro (1985), essa política pode ser entendida como um guia para a tomada de decisões, fundamental para determinar o tipo de serviço oferecido, para identificar os usuários e, consequentemente, para atender suas necessidades informacionais. Inclui também a definição dos recursos humanos, materiais e financeiros que delimitam o funcionamento de um sistema de recuperação da informação.

No processo de indexação estão envolvidos diversos aspectos que, mesmo formalizados, o caracterizam como uma atividade em que a subjetividade do indexador tem implicações profundas sobre a análise do documento. A subjetividade do indexador, aliada ao tempo gasto e ao custo alto são argumentos dos defensores da indexação automática (GIL LEIVA, 1999). Nos últimos anos se discute sobre a indexação automática tornar-se uma alternativa oportuna ao tratamento da informação. A análise do contexto de indexadores tem revelado a carga excessiva de trabalho enfrentado nas bibliotecas, motivo que, associado ao favorável avanço tecnológico, tem suscitado a expectativa em torno de alternativas para disponibilizar as informações de forma mais rápida e precisa.

Sem descartar as diferentes variáveis envolvidas na indexação realizada por humanos e na indexação automática, Moreiro González (2004) verifica que não se trata de justificar se é necessário ou não automatizar a indexação ou se o trabalho do indexador é mais ou menos custoso ou desnecessário. Trata-se de analisar que, nas atuais circunstâncias de crescimento informativo, a questão se centra na necessidade de criar um software eficaz que automatize o

[11] ZUNDE, P.; DEXTER, M. E. Indexing consistency and quality. *American Documentation*, p. 259-267, jul. 1969.

processo. Deve-se considerar que os documentos indexados de maneira automática respondem a padrões determinados e que a indexação automática não poderá dar conta de alguns aspectos que podem ser obtidos apenas por análise humana.

A aplicação da indexação automática tem-se desenvolvido como alternativa ao tratamento da informação diante do crescimento exponencial do volume de documentos. Essa circunstância é exposta por Robredo (2005) ao dizer que a necessidade de indexar grandes volumes de informações, em um tempo curto para manter as bases de dados atualizadas, tornou inviável pensar na indexação manual (humana ou intelectual) como única forma de analisar e codificar o conteúdo dos documentos. Dessa forma, Robredo (2005) defende que as pesquisas relacionadas à indexação automática devem-se desenvolver ao mesmo tempo em que as pesquisas em indexação manual (humana ou intelectual).

Diante de uma variedade de expressões apresentadas na literatura de Ciência da Informação, cerca de vinte expressões (GIL LEIVA, 1999) se referem à concepção de que a automatização da indexação compreende os conceitos que relacionam, de alguma forma, aplicação de sistema computacional à atividade de indexação — em realidade essas expressões dizem respeito a três conceitos: indexação assistida por computador, indexação semiautomática e indexação automática (GIL LEIVA, 1999; MOREIRO GONZÁLEZ, 2004).

A indexação assistida por computador refere-se ao processo em que o indexador humano realiza toda a atividade de análise do conteúdo do documento e utiliza um sistema computacional apenas para armazenar a representação da informação.

Já a indexação semiautomática está relacionada ao processo em que um sistema computacional realiza a atividade de análise do conteúdo do documento e, posteriormente, um indexador humano avalia os termos para indexação propostos pelo sistema.

Finalmente, no processo de indexação automática ocorre a atividade de análise do conteúdo do documento por um sistema computacional sem que haja uma avaliação posterior. Isto é, os termos de indexação são definidos apenas pela análise realizada pelo sistema (GIL LEIVA, 1999).

Nesse sentido, esta pesquisa pretende considerar os dois últimos conceitos, especialmente o conceito de indexação automática, visto que a análise comparativa do SISA aqui proposta será focalizada sobre seu processo automático, apesar de a proposta inicial do sistema ser semiautomática.

A análise de conteúdo do documento realizada na indexação automática constitui-se em um processo em que se aplicam métodos previamente estabelecidos, estando cada sistema de indexação automática sujeito, portanto, a aplicação de critérios estatísticos, linguísticos ou mistos. Verifica-se que a aplicação de tais métodos gera implicações sobre os resultados esperados na representação da informação e, consequentemente, na recuperação da informação. Assim, os principais aspectos teóricos e metodológicos, desde seu desenvolvimento inicial em meados do século XX até as propostas vigentes atualmente, são apresentados.

Diversos métodos de indexação automática foram desenvolvidos na tentativa de melhorar os resultados da indexação. No entanto, cada método possui avanços e também limitações em algum sentido, ou seja, um único método não satisfaz todas as exigências que garantam qualidade à atividade de indexação. Por isso, é importante destacar suas principais características, as relações de desenvolvimento entre os métodos e a relação de áreas do conhecimento no seu aprimoramento, no sentido de que, a partir dessa compreensão, será possível contextualizar nossa proposta.

Em meados do século XX, período pós-guerra, a produção científica e tecnológica foi impulsionada por grandes incentivos governamentais e privados. Áreas de pesquisa surgiram para atender a exigência cada vez maior de uma especialização do conhecimento científico.

É nesse cenário que a Ciência da Informação se origina, tendo, na realidade, impulsionado a consolidação de uma área que já tratava das questões de acesso informacional desde o final do século XIX, advindas principalmente da tradição biblioteconômica e da documentação. Assim ocorreu também com áreas ligadas ao desenvolvimento tecnológico, como a Ciência da Computação, que, de fato, se desenvolve a partir da exigência de tecnologias mais sofisticadas para a Segunda Guerra Mundial.

Nesse contexto, surgem as iniciativas de tratamento da informação com aplicação de sistemas computacionais. Os primeiros sistemas de indexação automática foram baseados exclusivamente em métodos estatísticos e probabilísticos, passando a incorporar métodos linguísticos somente a partir da década de 1980, ainda que esses estivessem desenvolvendo-se desde os anos 1960.

O princípio que norteou o desenvolvimento dos métodos estatísticos de indexação foi o “princípio do mínimo esforço”, proposto por Zipf em 1949. Segundo esse princípio, a razão constante entre a frequência das palavras e a posição que essas ocupam na ordem frequencial poderia indicar que a frequência das palavras em um texto tem relação com sua utilidade na

indexação, sugerindo, portanto, o critério de frequência para determinar se uma palavra seria considerada termo de indexação.

Hans Peter Luhn, em 1957, aplicou o princípio do mínimo esforço proposto por Zipf para distinguir os termos de indexação. No início, Luhn desenvolveu a indexação por meio da extração de palavras do título do documento para construção do índice *Key-Word In Context* (KWIC). Nesse tipo de índice, a palavra considerada como ponto de entrada é situada no centro, com o restante do título de ambos os lados (incluindo as palavras vazias[12]).

Esse método foi utilizado por Crestadoro na compilação do catálogo da Biblioteca Pública de Manchester, no século XIX; porém, o seu valor no processamento por computador foi estabelecido por Hans Peter Luhn (FOSKETT, 1973). O método de indexação automática empregado por Luhn na construção de índices consistia em confrontar uma lista de palavras vazias com o texto do documento e, dessa forma, eliminar as palavras insignificantes, tais como artigos, preposições e conjunções, restando, assim, as palavras que figurariam como termos de indexação FIG.4.

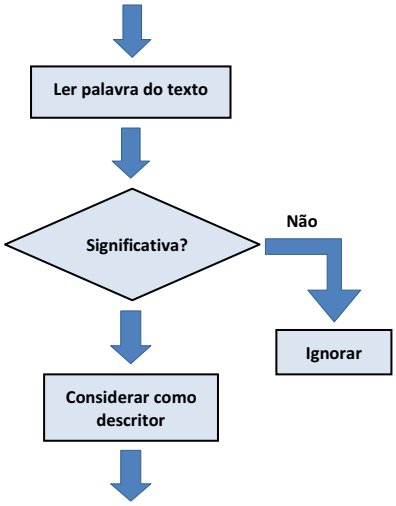


FIGURA 4 - Algoritmo básico do processo de indexação automática
Fonte: Robredo (1991)

Apesar de haver redução das palavras do texto pelo emprego da lista de palavras vazias, esse método, muito simples, gerava muitas entradas temáticas no índice, o que exigia a aplicação de outra forma de filtro após a eliminação de palavras vazias. Nesse sentido, o critério de frequência baseado no princípio de Zipf foi aplicado para determinar os termos de indexação.

[12] Conhecidas também por "palavras proibidas", ou "*stopwords*", em inglês, são palavras irrelevantes para indexação, tais como artigos, preposições, conjunções, etc.

Estudos de indexação ponderada derivaram da concepção de frequência como critério, atribuindo-se valor ou pesos de importância aos termos de indexação. A frequência relativa é calculada a partir da ocorrência de palavras nos documentos. A palavra é atribuída como termo de indexação com relação à sua capacidade de distinguir os documentos de uma coleção, como, por exemplo, a função de frequência inversa (*inverse document frequency weight, IDF*) proposta por Sparck Jones[13] (1972) e os métodos de valor de discriminação dos termos, proposto por Salton e Yang[14] (1973) (VIEIRA, 1988; GIL LEIVA, 1996, 1999, 2008; MENDEZ RODRÍGUEZ e MOREIRO GONZÁLEZ, 1999; MOREIRO GONZÁLEZ, 2004).

A função de frequência inversa de um documento examina a ocorrência de um termo na coleção de documentos, considerando que a frequência com que um termo aparece está em relação inversa à sua capacidade informativa (GIL LEIVA, 2008).

O valor de discriminação dos termos é um método para determinar o valor daqueles termos que têm a capacidade de distinguir os documentos da coleção (GIL LEIVA, 2008). Por meio dessa medida é possível identificar os termos que são bons discriminantes, ou seja, aqueles que permitem oferecer um diferencial com relação a outros termos que representam os documentos da coleção, proporcionando mais precisão na busca de documentos.

Mendez Rodríguez e Moreiro González (1999) ressaltam a importância significativa dos primeiros modelos de indexação automática baseados em critérios estatísticos ou probabilísticos, pois foram os primeiros métodos que surgiram como alternativa à indexação automática, aproveitando o avanço da informática. Ainda continuam sendo aplicados integrados aos métodos linguísticos para indexação, assim como para extração de palavras nos processos de elaboração de linguagens controladas, como os tesauros.

Os primeiros métodos de indexação automática fundamentaram-se na análise estatística e probabilística do texto, tomando principalmente a palavra como objeto de análise para identificação de termos relevantes para indexação (MENDEZ RODRÍGUEZ, MOREIRO GONZÁLEZ, 1999).

Houve dificuldades, pois a aplicação apenas de métodos estatísticos não é capaz de distinguir as variações linguísticas dos termos da linguagem natural, suscetíveis a fenômenos como sinonímias, polissemias, homonímias, anáforas, elipses, formas flexionadas de gênero e número, termos constituídos por mais de uma unidade lexical, termos apresentados em

[13] SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, v. 28, p. 11-21, 1972.
[14] SALTON, G.; YANG, C. S. On the specification of term values in automatic indexing. *Journal of Documentation*, v. 29, n. 4, p. 351-372, 1973.

formato de siglas ou sua forma por extenso, conceitos implícitos e outras situações da linguagem.

Diante dessas dificuldades, métodos linguísticos de tratamento automático foram desenvolvidos, principalmente a partir dos estudos de Processamento de Linguagem Natural (PLN) fundamentados na área de Linguística Computacional. Segundo Othero e Menuzzi (2005, p. 25), a Linguística Computacional "é uma área da ciência linguística voltada para o tratamento computacional da linguagem e das línguas naturais"; seu desenvolvimento se deu principalmente a partir dos estudos de tradução automática nos anos 1950 e 1960.

Os sistemas de indexação automática com base em métodos linguísticos foram desenvolvidos a partir da década de 1960, integrando analisadores linguísticos dedicados a solucionar dificuldades de análise morfológica, sintática e semântica.

Os analisadores morfológicos dedicam-se à análise de palavras que constituem os textos, buscando identificar os elementos que as constituem. Por exemplo, as raízes e afixos (prefixos e sufixos) e os fenômenos de flexão e derivação que estão relacionados à sua composição.

A análise linguística inicia-se por uma análise tipográfica e ortográfica em que ocorre a segmentação do texto em frases, para identificação de erros tipográficos, de ortografia e de acentuação. A partir daí, atua a análise morfológica, que reconhece as palavras considerando as formas flexionadas. E, por fim, a análise léxica, que reduz as variações dessas palavras em sua forma ou entrada de dicionários (infinitivo para verbos, masculino singular ou feminino singular para nomes, e masculino singular para os adjetivos) (MOREIRO GONZÁLEZ, 2004).

Desse modo, é possível atuar por meio de um processo de lematização. Nesse processo ocorre a redução de uma palavra ou conjunto de palavras à sua raiz, uma vez detectadas ou eliminadas suas formas flexivas (número, gênero, desinência) e derivativas (-ístico; -ável; -dade; -ista; -ção, etc.) mediante um sistema computacional, para que se possa calcular a frequência da ocorrência de um termo a partir de uma mesma raiz identificada (GIL LEIVA, 2008). Esse procedimento pode ser útil para unificar o tratamento estatístico e facilitar as operações de filtro, de criação de relações e de redes semânticas de representação dos conceitos (MOREIRO GONZÁLEZ, 2004).

O procedimento de lematização, além de ser aplicado no tratamento da informação, pode ser aplicado à recuperação da informação, uma vez que é possível conectar um termo de busca a todos os termos com a mesma raiz. A lematização gera índices mais concisos e

aumenta a revocação, pois multiplica os pontos de acesso a determinados documentos, na medida em que um único morfema pode estar associado a muitas palavras diferentes (SOUZA, 2005).

Contudo, é necessário lembrar que realizar esse processo automaticamente torna-se complicado em algumas circunstâncias, pois existem regras estabelecidas não amplamente aplicáveis — por exemplo, quando os "s" são eliminados, na tentativa de excluir os plurais dos índices, o que não se aplica a palavras com plural diferente e que quebram essa regra (CÂMARA JUNIOR, 2007).

Outro aspecto a ser considerado é o conceito desses termos, visto que os termos podem apresentar a mesma raiz mas não possuir o mesmo significado. Por exemplo, os termos "***Index****ação*", "***Index****ador*" e "***Index****ável*", que representam o processo, o agente e um atributo, respectivamente. Pode ser problemático considerá-los como um mesmo conceito (ANDERSON & PÉREZ-CARBALLO, 2001).

Outro processo que atua nesse nível de análise é conhecido como "tokenização", entendido como o procedimento em que ocorre a separação do texto em palavras através do reconhecimento do texto entre determinadas marcas (CÂMARA JUNIOR, 2007).

É necessário definir as marcas que não são interessantes como descritores, como os espaços em branco, hífens e sinais de pontuação que delimitam um token. Porém, é preciso analisar cuidadosamente essa decisão. Existem situações em que a tokenização pode ser necessária, e, por outro lado, situações em que deve ser ignorada, para preservar o significado (CÂMARA JUNIIOR, 2007).

Anderson & Pérez-Carballo (2001) também têm exposto o problema de considerar pontuações e marcas, como os hífens, para delimitar palavras, uma vez que esses recursos podem indicar a conexão de palavras, assim como palavras isoladas. Uma solução indicada pelos autores é considerar todas as possíveis combinações de uma palavra.

Outra questão exposta por Anderson e Pérez Carballo (2001) é a indexação automática de números, considerando que os números possuem diferentes funções segundo o tipo de texto. Questão similar acontece com relação à identificação de palavras com apenas um caractere, como na locução "vitamina C", em que o caractere "C" possui um papel importante na garantia de significado e, portanto, não deve ser suprimido por um processo de eliminação de palavras vazias.

Cabe mencionar que o procedimento de eliminação de palavras vazias também atua nesse nível de análise. Realiza um processo de filtro, eliminando as palavras insignificantes para indexação. Para elaborar essa lista, é preciso verificar cuidadosamente se a eliminação de uma palavra comprometerá a indexação, pois pode constar como parte de um termo constituído por mais de uma unidade lexical ou pode a sua relevância estar sujeita à área de conhecimento em que será aplicada (MOREIRO GONZÁLEZ, 2004).

Apresentamos a seguir um exemplo do processo de análise morfológica em uma frase simples como "a planta da casa estava na mesa". É possível verificar sua decomposição pelo analisador morfológico nas seguintes partes:

QUADRO 7 - Análise morfológica de uma frase

A planta da casa estava na mesa			
Expressão	**Categoria**	**Lema**	**Informação morfológica**
A	Preposição	a	
	Artigo	o	Singular, Definido, Feminino
	Pronome pessoal	o	Singular, Terceira pessoa, Feminino
planta	Substantivo comum: Mapa/Vegetal	planta	Singular, Feminino
	Verbo	plantar	Presente, Terceira pessoa, Singular, Transitivo
	Verbo	plantar	Imperativo, Segunda pessoa, Singular, Transitivo
da	Contração da preposição "de" com o artigo definido "a"	do	Singular, Feminino
casa	Substantivo comum	casa	Singular, Feminino
	Verbo	casar	Presente, Singular, Terceira pessoa, Transitivo
	Verbo	casar	Imperativo, Singular, Segunda pessoa, Transitivo
estava	Verbo	estar	Pretérito imperfeito, Singular, Primeira pessoa, Transitivo/Intransitivo
	Verbo	estar	Pretérito imperfeito, Singular, Terceira pessoa, Transitivo/Intransitivo
na	Contração da preposição "em" com o artigo definido "a"	no	Singular, Feminino
mesa	Substantivo comum	mesa	Singular, Feminino

Fonte: Elaborado pela autora com base no analisador morfológico Webjspell[15]

[15] Analisador morfológico disponível em: <http://natura.di.uminho.pt/webjspell/jsol.pl>

Desse modo, podemos dizer que a importância do nível de análise morfológica se justifica por permitir uma análise profunda da estrutura e da formação das palavras. Embora essa análise não contemple a identificação especificamente dos aspectos semânticos envolvidos no tratamento da informação, constitui alicerce para que o próximo nível possa ser realizado pelos analisadores sintáticos.

No exemplo acima, é possível verificar que o analisador morfológico apresenta as possíveis categorias gramaticais de cada elemento da frase, mas sem definir a categoria correta no contexto da frase. Verificamos, por exemplo, a palavra “planta”, que, de acordo com o analisador morfológico, pode ser um verbo ou um substantivo, e que este, ainda, pode ter o significado de “mapa” ou de “vegetal”. Nesse caso, o analisador sintático atuará para resolver a ambiguidade.

Para tanto, os analisadores sintáticos atuam por meio de uma “gramática”, que consiste, de fato, em um dicionário com palavras e suas possíveis categorias gramaticais formalizadas na análise morfológica. Um algoritmo de análise soluciona a ambiguidade e define as relações entre as palavras. Além disso, determina a disposição das palavras nas orações, sua função e a combinação entre as palavras para obter orações gramaticalmente corretas (MOREIRO GONZÁLEZ, 2004; GIL LEIVA, 2008).

Nesse sentido, verifica-se que o analisador sintático opera sobre os problemas relativos à ambiguidade das categorias gramaticais assinaladas pelo analisador morfológico, já que este apenas apresenta as possíveis categorias gramaticais, o gênero e o número de cada palavra sem defini-las no contexto em que a palavra aparece, o que apenas sucede na análise sintática (GIL LEIVA, 2008).

Em síntese, na análise sintática acontece o procedimento em que as palavras são definidas em função de seu papel, no contexto em que ocorrem em uma oração, determinadas segundo princípios de construção e coordenação das frases, disposição das palavras na oração, e das orações no período. A relação dessa análise, assim como da análise morfológica, com a análise semântica, define-se pela elaboração e estabelecimento de dados formalizados, para que o analisador semântico identifique o significado expresso no texto do documento.

É importante destacar que, por mais que esses dados estejam formalizados, o nível de análise semântica almejado torna-se de difícil alcance, visto que, na busca do significado das orações, estão muitas vezes envolvidos aspectos implícitos, muito além das estruturas e formalismos explícitos demarcados nas análises morfológicas e sintáticas.

A análise semântica objetiva descobrir o significado das palavras (semântica léxica), reconhecer sinônimos, situar o significado das palavras dentro das orações (semântica gramatical), estabelecer o conjunto de palavras que se relacionam com um mesmo campo semântico (semântica contextual), determinar os termos gerais e específicos e estabelecer enlaces com os antônimos (MOREIRO GONZÁLEZ, 2004).Para que essas condições possam realmente ser concretizadas, é necessário amplo conhecimento sobre as palavras e seu significado no universo do discurso, com o objetivo de formalizar tais interpretações (GIL LEIVA, 1999), exigindo, muitas vezes, a associação de instrumentos linguísticos e terminológicos, como bases de conhecimento, ontologias, tesauros, vocabulários controlados, listas de descritores, etc.

Dessa forma, o que se espera dos analisadores semânticos é que operem mediante processos de inferências para extraírem conhecimento do conteúdo dos documentos e representá-los na forma de termos. Ou seja, que tenham a capacidade de reconhecer conceitos, identificando o significado das palavras e orações, considerando fenômenos como as sinonímias, as anáforas, frases e palavras compostas, homógrafos, homonímias, as polissemias, e introduzindo relações de hierarquia entre as palavras (MOREIRO GONZÁLEZ, 2004).

Em contraposição à iniciativa da indexação automática, derivada exclusivamente de métodos estatísticos ou apenas de métodos linguísticos, as propostas atuais são de integração de métodos.

Com base no modelo de arquitetura de um sistema de indexação automática, apresentado por Gil Leiva (2008), constata-se a atuação da análise linguística nos tratamentos morfológico, sintático e semântico. Em seguida, opera o analisador estatístico, que efetua os cálculos de frequência de ocorrência sobre dados tratados linguisticamente e formalizados em etapa anterior. Subsequentemente aplica-se um vocabulário controlado, em que este é cotejado com os termos candidatos à indexação, permitindo a atribuição definitiva dos termos de indexação (FIG. 5).

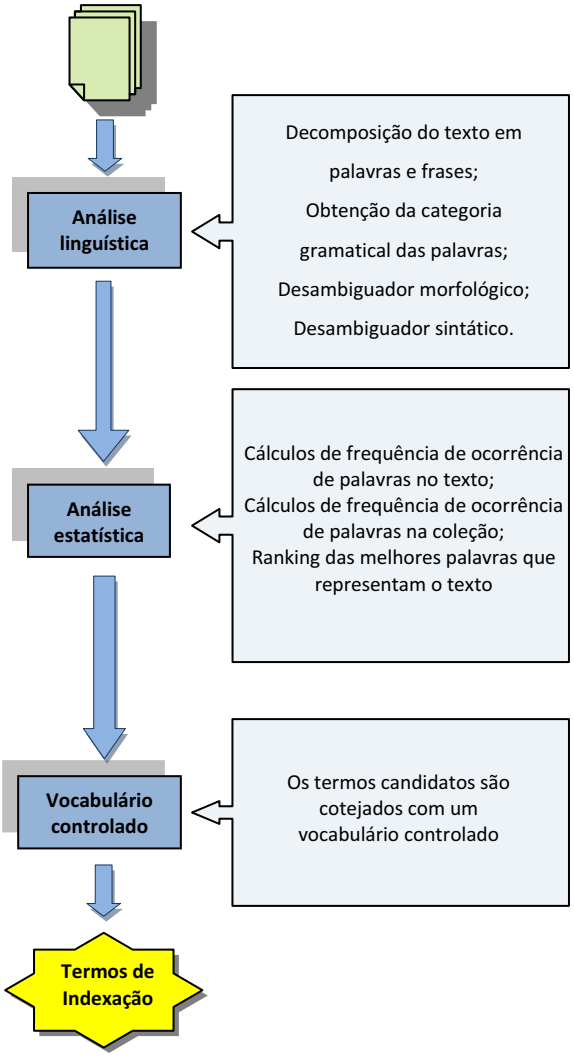


FIGURA 5 - Modelo de arquitetura de um sistema de indexação automática
Fonte: GIL LEIVA, 2008, p. 367.

Com exceção dos primeiros sistemas com métodos puramente estatísticos, verificam-se os métodos mistos ou híbridos de indexação automática, que reúnem aportes da estatística e da linguística textual e ainda utilizam tesauros como instrumento de controle de vocabulário, contribuindo para eliminar problemas como a sinonímia e a identificação de funções sintáticas dos termos, proporcionando benefícios à revocação na recuperação da informação (GIL LEIVA, 1999; GUIMARÃES, 2000).

A aplicação de vocabulário controlado em indexação automática permite definir o significado das palavras, seleciona as candidatas à indexação no contexto de uma área de conhecimento e auxilia na determinação dos termos que, de fato, serão atribuídos para indexação.

Com relação ao modo com que os termos de indexação são selecionados, Lancaster (2004) distingue dois tipos de indexação: a *indexação por extração automática* e a *indexação por atribuição automática*.

Na indexação por extração automática, as palavras ou expressões que aparecem no texto são extraídas e utilizadas para representar o texto como um todo, ou seja, a indexação é realizada a partir da linguagem natural. Já a indexação por atribuição automática consiste na representação do conteúdo mediante termos selecionados de alguma linguagem de indexação. Segundo Lancaster (2004), a indexação com uso de linguagens de indexação é realizada na maior parte das indexações por seres humanos e é considerada mais difícil de ser aplicada por computadores.

Como descrito no esquema apresentado, em muitos sistemas de indexação automática ocorre a indexação por atribuição automática. Sucede um processo em que o vocabulário controlado é confrontado com os termos candidatos à indexação, selecionados por critérios estatísticos e linguísticos. É um processo delicado, muitas vezes, devido a essa atribuição dos termos identificados no vocabulário controlado ocorrer por meio da identificação de padrões, isto é, por meio da identificação de sequência de caracteres comuns. Isso pode implicar em conflito de significados que compromete a qualidade dos resultados da indexação automática.

Portanto, dessa abordagem pudemos constatar que existem três conceitos na literatura que se referem ao processo de automatizar a indexação: “indexação assistida por computador”, “indexação semiautomática” e “indexação automática”, em que cada conceito caracteriza processos distintos.

Os primeiros sistemas de indexação automática se fundamentaram em métodos estatísticos e probabilísticos, passando depois a incorporar métodos de base linguística — estes, constituídos por analisadores morfológicos, sintáticos e semânticos, em que cada analisador cumpre a função de oferecer subsídios para o analisador seguinte. Verifica-se, entretanto dificuldades ocasionadas por limitações nas ferramentas linguísticas de cada analisador.

Como verificamos, recursos linguísticos como a lematização e a tokenização oferecem uma alternativa para solucionar algumas situações que ocorrem na linguagem, mas existem exceções não contempladas e que devem ser cuidadosamente analisadas. Tal fato é explicado pela natureza dinâmica da linguagem.

É por isso que se torna complexo aplicar um vocabulário controlado automaticamente, atribuindo ao sistema a tarefa de analisar e propor os termos que melhor representam o

conteúdo temático do documento. A decisão sobre os melhores descritores que representam o documento é uma atitude de reflexão bastante complexa para o indexador humano, reflexão impossível na indexação automática, que se baseia em critérios objetivos, mas que precisam ser um tanto flexíveis para contemplar diversas situações.

Nesse contexto, também estão envolvidas as diretrizes que orientam a realização dessa indexação. Isto é, os aspectos que constituem a política de indexação apontam a forma com que o sistema de indexação automática deverá atuar para atender as necessidades e particularidades de determinado sistema de informação. Uma característica importante a considerar é o domínio do conhecimento em que o sistema de indexação automática atuará, uma vez que as particularidades da literatura, das fontes de informação e da terminologia de cada área podem influir no desempenho do sistema e merecem análise e adaptação.

No capítulo seguinte apresentamos os procedimentos metodológicos empregados para realizar a revisão teórica sobre os sistemas de indexação automática, assim como para realizar o experimento de aplicação e avaliação do ThesAgro no sistema SISA.

4 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo expomos os procedimentos empregados na pesquisa que compreendem a sistematização dos aportes teóricos e análise das propostas e sistemas de indexação automática, a aplicação do vocabulário ThesAgro no sistema SISA e a avaliação da indexação. Tais procedimentos se integram na medida em que os aportes teóricos sobre vocabulários controlados, indexação automática e sistemas de indexação embasam as discussões que se originaram da análise dos resultados de aplicação e avaliação da indexação. A aplicação do ThesAgro no sistema SISA fornece os dados quantitativos e os descritores que são examinados na fase de avaliação (capítulo 5). Nesta fase, por sua vez, são obtidos os fatores que interferiram na indexação e recuperação da informação (capítulo 6). Por fim, os resultados da avaliação em consonância com os resultados teóricos é que permitem sintetizar os aspectos a serem considerados na adaptação de vocabulários controlados no processo de indexação automática (capítulo 7).

4.1 Sistematização teórica sobre indexação automática e sistemas de indexação automática

Esta pesquisa tem caráter teórico-prático, sendo desenvolvida por abordagem qualitativa (revisão da literatura) e por abordagem quantitativa (investigação dos resultados propostos pelo experimento com o SISA). Nesse sentido, para a elaboração da revisão da literatura realizamos uma pesquisa exploratória com o objetivo de "proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses" (GIL, 1996, p.45).

Para tanto, desenvolvemos uma pesquisa bibliográfica da literatura nacional e internacional da área de Ciência da Informação e áreas afins. O levantamento bibliográfico foi realizado a partir da sistematização de pesquisas, valendo-se das palavras-chave "*indexação automática*", "*linguagens de indexação*", "*linguagens documentárias*", "*tesauros*", "*cabeçalhos de assuntos*", "*vocabulário controlado*", "*sistemas de indexação automática", "software de indexação*", além de suas variações e da tradução em inglês, considerando as pesquisas publicadas entre o período de 1950 a 2011.

As estratégias de busca foram realizadas nas bases de dados: Athena (base de dados bibliográfica da Universidade Estadual Paulista); Dedalus (base de dados bibliográfica da

Universidade de São Paulo); Biblioteca Digital Brasileira de Teses e Dissertações (BDTD); Library Information Science Abstract (LISA); Library, Information Science and Technology Abstracts (LISTA); Wilson Web; Emerald; Portal de Periódicos Capes; Portal de Periódicos Scielo; além de buscas em periódicos nacionais *on-line* como "DataGramazero", "Ciência da Informação", "Perspectivas em Ciência da Informação" e "Informação & Sociedade: Estudos".

A partir do levantamento bibliográfico e da leitura do título, do resumo e de trechos significativos das obras, selecionamos as fontes de informação que abordam especificamente os aspectos históricos, conceituais e metodológicos da indexação automática, tipos de linguagens de indexação como os cabeçalhos de assuntos e os tesauros, e também as ontologias. Também buscamos fontes de informação sobre os sistemas de indexação automática que aplicam alguma linguagem de indexação, PLN ou ontologia. A seleção e a análise das fontes de informação para a elaboração dos capítulos teóricos esteve pautada, sobretudo, nos aspectos considerados relevantes para a análise dos dados da pesquisa.

A sistematização e a leitura dos referenciais teóricos nos permitiram desenvolver três capítulos teóricos. O capítulo "*Os vocabulários controlados como linguagens de indexação*" apresenta um panorama do desenvolvimento das linguagens de indexação, desde os princípios de Cutter para a elaboração de cabeçalhos de assuntos aos tesauros, e, mais recentemente, a aplicação de ontologias. A intenção não foi apresentar um referencial exaustivo, mas um panorama do desenvolvimento e das características das linguagens de indexação, para guiar a análise da pesquisa, em que se pressupõe que os aspectos vigentes nos atuais vocabulários controlados — sob influência das correntes que os originaram — aplicados na pesquisa podem influenciar os resultados de indexação obtidos na indexação automática do SISA.

Em seguida, foi desenvolvido o capítulo "*Indexação Automática*", em que tratamos especificamente dos aspectos conceituais, históricos e metodológicos da indexação automática. O foco do capítulo é o modo como são realizados os processos de indexação automática e seu aperfeiçoamento desde meados do século passado até os dias atuais. Esses aspectos metodológicos se associam às características dos sistemas de indexação descritas e analisadas no capítulo seguinte.

Nesse sentido, foram analisados 20 sistemas de indexação automática apresentados no capítulo *"Sistemas de indexação automática"*. A escolha dos sistemas e sua apresentação foram conduzidas por critérios de importância histórica, proposta metodológica e uso de vocabulário controlado em indexação por atribuição.

O critério de importância histórica contribui para a contextualização dos processos de indexação automática empregados no início do desenvolvimento da área, proporcionando uma análise sobre as mudanças que se fizeram necessárias diante de suas limitações. Por outro lado, o critério da proposta metodológica nos fornece informações sobre as alternativas investigadas para solucionar determinados problemas enfrentados nos sistemas de indexação automática. O último critério contribui diretamente para a análise dos resultados da pesquisa ao apresentar a aplicação de vocabulários controlados em sistemas de indexação automática, permitindo a análise sobre as implicações no processo de indexação.

QUADRO 8 - Critérios para seleção de sistemas de indexação automática

Critério	Sistemas de indexação
Importância histórica	KWIC, KWOC e KWAC PRECIS POPSI NEPHIS e LIPHIS
Proposta metodológica	SMART (identificação de termos compostos) Zstation (solução de ambiguidades) Sintagmas Nominais (KURAMOTO, 2002) (identificação de sintagmas nominais) Proposta da UTC (identificação de Unidades Terminológicas Complexas) Sintagmas Nominais (SOUZA, 2005) (identificação de sintagmas nominais) SiRILiCO (análise sintática e semântica) Indexação de acórdãos (CÂMARA JÚNIOR, 2007) (indexação automática de acórdãos) Algoritmos genéticos (representação dos documentos adaptada às necessidades dos usuários) SintagMed (indexação automática de laudos médicos)
Uso de vocabulário controlado em indexação por atribuição	FAIRS AUTOMINDEX Concept Indexer HEPIndexer AUTINDEX Sistema multilíngue (POULIQUEN, STEINBERGER e IGNAT, 2003) CADIS

Fonte: Elaborado pela autora

O QUADRO 8 apresenta o critério principal de seleção da proposta ou sistema de indexação para análise na pesquisa. Isso significa que os sistemas também poderiam ser selecionados por outros critérios — por exemplo, o sistema AUTOMINDEX, que possui importância histórica no desenvolvimento de sistemas de indexação automática no Brasil —, mas foram selecionados pelo critério que mais se destacou para a necessidade da pesquisa.

A descrição de cada sistema de indexação automática e a análise de suas características esteve apoiada pela leitura e pela interpretação das publicações disponíveis, esclarecendo que, em nenhum momento, tivemos acesso aos sistemas propriamente ditos, o que não impossibilitou a análise de suas principais características.

O referencial teórico foi analisado sob perspectiva qualitativa, recorrendo à interpretação e à reflexão sobre os aspectos caracterizadores do tema abordado. Esses aspectos foram resgatados na etapa de análise dos resultados da pesquisa, em que verificamos as origens dos problemas levantados na atuação do vocabulário controlado na indexação automática realizada pelo SISA e também os associamos aos problemas constatados em pesquisa anterior de Narukawa, Gil Leiva e Fujita (2009).

4.2 Metodologia de aplicação do SISA com uso de vocabulário controlado

Aplicamos o SISA pela possibilidade de acesso ao sistema que nos foi oferecida na parceria de pesquisa firmada entre a Professora Doutora Mariângela Spotti Lopes Fujita, da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), e o pesquisador Professor Doutor Isidoro Gil Leiva, da Universidade de Murcia, da Espanha.

O SISA é um sistema de indexação desenvolvido pelo Professor Doutor Isidoro Gil Leiva durante seu doutorado, concluído em 1999. A proposta é aplicá-lo a artigos científicos, haja vista as estruturas do documento analisadas pelo sistema, delimitadas por marcadores específicos que indicam o título (#CTI# e #FTI#), o resumo (#CR# e #FR#) e o texto (#CTE# e #FTE#). O documento em texto completo é uma das fontes exigidas pelo SISA, além de um vocabulário controlado e uma lista de palavras vazias.

O processo de análise automática ocorre em três módulos:

Módulo I: fase de pré-processamento, em que o documento inserido no sistema é inicialmente sinalizado com os marcadores correspondentes a título, resumo e texto, para que, posteriormente, os cálculos de ponderação sejam realizados a partir da identificação da frequência nessas estruturas.

Ocorre também o processo de eliminação das palavras vazias por meio do confronto do documento com a lista de palavras vazias.

Além disso, ocorre a etapa de horizontalização, em que frases e orações compreendidas entre os sinais de pontuação (. , ; :) são dispostas em forma horizontal, ou seja, são separadas em cada linha do texto.

Módulo II: fase de processamento, em que o documento é analisado e se buscam e selecionam-se os termos de indexação. Na FIG. 6 é possível verificar em detalhes como ocorre o processo (GIL LEIVA, 2008):

1º) Extrai-se o primeiro termo do vocabulário controlado;

2º) Extrai-se o primeiro termo da fonte[16];

3º) Verifica-se se os termos extraídos são iguais;

4º) Se não são iguais, verifica-se se existem mais termos nas fontes. Se não existem mais palavras, verifica-se se existem mais termos no vocabulário controlado. Se não houver, o processo é finalizado. No caso de haver mais palavras em alguma fonte, extrai-se a seguinte, que é novamente comparada com o termo do vocabulário controlado. Esse processo se repete até serem iguais ou não existirem mais palavras nas fontes.

5º) Se a palavra da fonte e o termo do vocabulário controlado são iguais, verifica-se se a entrada do vocabulário controlado tem subentradas:

- Se não tem subentrada, verifica-se se há relação de equivalência da entrada principal, sendo o termo autorizado[17] enviado ao módulo de candidato a termo de indexação. Esse termo é marcado para que não seja necessário localizá-lo novamente.
- Se tiver subentrada, verifica-se se coincide com alguma das quatro palavras seguintes à direita da última palavra localizada na fonte.

6º) Se alguma das quatro palavras coincide com o termo do vocabulário controlado, comprova-se se este tem subentradas e repete-se o mesmo processo. No caso de não ter mais subentradas ou nenhuma das quatro palavras da fonte coincidir com a do vocabulário, comprova-se se existem relações de equivalência e transfere-se (transferem-se) o(s) termo(s) autorizado(s) ao módulo de candidatos e marca-se (marcam-se) o(s) localizado(s).

[16] "Fonte" se refere às estruturas constituintes do documento demarcadas em título, resumo e texto.
[17] "Termo autorizado" se refere ao termo que consta no vocabulário controlado e que permite utilizá-lo como termo de indexação.

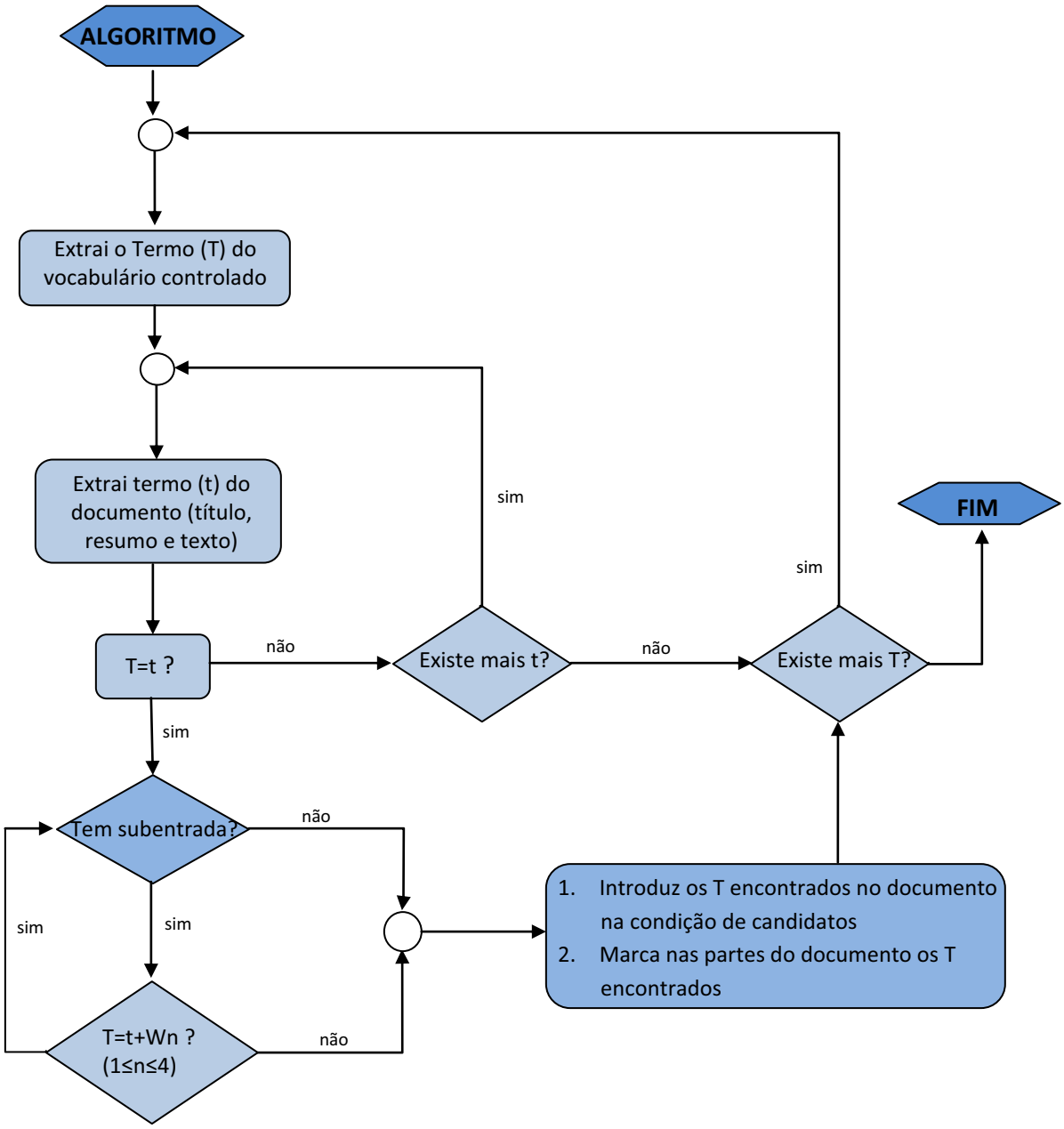


FIGURA 6 - Diagrama de fluxos do algoritmo SISA
Fonte: Gil Leiva (2008, p.380)

7º) Confirma-se se há mais termos no vocabulário controlado:

- Se não existem mais palavras, finaliza-se o processo.
- Se houver mais palavras, continua-se o processo desde a 2ª fase.

Módulo III: fase de ponderação dos termos que foram analisados pelo sistema; do contrário, todos os termos do vocabulário controlado que coincidem com os das fontes do documento seriam selecionados e propostos como termos de indexação.

Nesse sentido, o sistema considera os seguintes critérios para propor os termos de indexação:

- Se um termo autorizado aparece na fonte "título" e na fonte "resumo", apresenta-se como termo de indexação.
- Se um termo autorizado aparece na fonte "título" e na fonte "texto", apresenta-se como termo de indexação.
- Se um termo autorizado aparece na fonte "resumo" e na fonte "texto", apresenta-se como termo de indexação.
- Se o termo candidato a descritor[18] aparece no título, no resumo e no texto, apresenta-se ao indexador para sua possível incorporação como termo de indexação.
- Se um termo candidato a descritor aparece no texto dez vezes ou mais, além de aparecer em oito parágrafos diferentes ou mais, e não está incluído em nenhum dos termos propostos, apresenta-se como termo candidato a termo de indexação.

O SISA foi proposto originalmente como um sistema de indexação semiautomático[19] aplicado à área de Biblioteconomia e Documentação, mas permite que seja aplicado a qualquer área do conhecimento, com a condição de que os arquivos de configuração sejam adaptados.

Para atender às condições de adaptação do sistema para o experimento, definimos a área de produção científica agrícola ao considerar a possibilidade de se obter um vocabulário controlado e artigos científicos em texto completo. Além disso, os termos de indexação dos artigos científicos estão disponíveis para consulta *on-line* na base de dados Base Bibliográfica da Agricultura Brasileira (AGROBASE), o que permite compará-los com os termos de indexação obtidos na indexação automática do SISA. Ademais, buscamos definir uma área do conhecimento distinta da área de odontologia e que possuísse vocabulário controlado para que pudéssemos confrontar os resultados do estudo com o resultados alcançados na aplicação do SISA na área de odontologia.

A BINAGRI foi criada como agente do Sistema Nacional de Informação Agrícola (SNIDA), com a finalidade de coletar, processar, armazenar e disseminar informações

[18] "Termo candidato a descritor" se refere ao termo que não foi localizado no vocabulário controlado, mas apresenta elevada ocorrência no documento.
[19] Para os objetivos desta pesquisa, considera-se somente o processo de indexação automática do SISA, ou seja, sem a validação por indexadores humanos.

científicas e tecnológicas de interesse do setor agrícola e áreas correlatas, garantindo a preservação da Memória Agrícola Nacional (BIBLIOTECA NACIONAL DE AGRICULTURA...).

A BINAGRI é uma Coordenação Geral de Informação Documental que dá suporte informacional a todos os órgãos do Ministério da Agricultura, Pecuária e Abastecimento (MAPA), bem como divulga os serviços de todo o Ministério, além de ser o órgão que gerencia a comercialização das publicações do MAPA. Em âmbito internacional, a BINAGRI mantém estreito relacionamento com sistemas e redes internacionais de informação documental agrícola, como o Sistema Internacional de Informação para a Ciência e Tecnologia Agrícola (Agris) da *Food and Agriculture Organization of the United Nations* (FAO), enviando regularmente os registros bibliográficos nacionais. Mantém parceira também com a Rede Internacional de Bibliotecas Agrícolas (AGLINET) e com o *Sistema de Información y Documentación Agropecuario de América* (SIDALC/AGRI2000), com a disponibilização da base de dados e da literatura nacional AGROBASE. Além disso, a BINAGRI mantém no acervo coleções de publicações dos seguintes órgãos: Instituto Interamericano de Cooperação para a Agricultura (IICA), FAO, *United State Department of Agriculture* (USDA/NAL) *e International Center for Tropical Agriculture* (CIAT) (BIBLIOTECA NACIONAL DE AGRICULTURA...).

Em âmbito nacional, faz o intercâmbio com instituições agrícolas brasileiras, disponibilizando documentos através de listas contendo referências bibliográficas das duplicatas recebidas, além de disponibilizar a todos os usuários que trabalham com informação um instrumento normalizador da terminologia agrícola brasileira, o ThesAgro (BIBLIOTECA NACIONAL DE AGRICULTURA...).

O ThesAgro é um tesauro especializado na literatura agrícola, aplicado à indexação e recuperação de documentos, tendo sido desenvolvido pela BINAGRI (BIBLIOTECA NACIONAL DE AGRICULTURA...).

Foi desenvolvido de acordo com as diretrizes da UNESCO, das normas do *Principles directeurs pour Létablissement et le développement the thesaurus monolingues* (SC/WS/555, Paris, 1973), tendo sido lançada, a sua primeira versão, em 1979. Uma versão enriquecida e melhorada foi lançada em 1989 e, atualmente, contém 9.351 termos, com uma versão disponível na internet para que outras instituições agrícolas brasileiras possam utilizar e colaborar no aperfeiçoamento do tesauro (BIBLIOTECA NACIONAL DE AGRICULTURA...).

Os 100 artigos científicos da área agrícola selecionados para a pesquisa são publicações dos fascículos: volume 28, números 1 e 2, de 2006, e volume 29, número 1, de 2007, da Revista Brasileira de Fruticultura (ISSN 1806-9967). Esta publicação foi criada em 1978, junto à Sociedade Brasileira de Fruticultura, e destina-se à divulgação de artigos técnico-científicos e comunicações científicas na área de fruticultura. Tal escolha se deve ao fato dos artigos científicos apresentarem os requisitos de formatação que o SISA exige para a execução da indexação automática, contendo título, resumo e texto completo em formato padronizado.

Nesse sentido, inicialmente realizamos a configuração dos arquivos utilizados no SISA, os testes de indexação em alguns artigos científicos e a indexação definitiva dos 100 artigos selecionados para, em seguida, proceder à avaliação da indexação com relação à análise de consistência na indexação e à análise da recuperação de informação.

4.2.1 Preparação dos arquivos utilizados no SISA

O SISA exige, em sua configuração, a entrada de três tipos de arquivos: uma linguagem de indexação, uma lista de palavras vazias e os artigos científicos. Para efetuar a indexação automática, é necessário preparar os arquivos utilizados pelo SISA em formato TXT, sendo que cada arquivo exige uma configuração específica:

a) **Linguagem de indexação:** Uma lista de descritores foi elaborada a partir do ThesAgro. Essa lista é organizada em ordem alfabética, enumerada, e os descritores estabelecem apenas a relação de equivalência por meio da indicação USE (Apêndice A).

b) **Lista de palavras vazias:** A lista de palavras vazias em língua portuguesa foi adaptada, a partir da conferência das palavras quanto ao fato de fazerem ou não parte da lista de descritores configurada no SISA, uma vez que a sua presença interfere na atribuição de termos de indexação (Apêndice B).

c) **Artigos científicos:** Selecionamos 100 artigos científicos agrícolas da base de dados AGROBASE. O título, o resumo e o texto, sem as referências dos artigos científicos, foram delimitados por marcadores #CTI# (começo do título), #FTI# (fim do título), #CR# (começo do resumo), #FR# (fim do resumo), #CTE# (começo do texto) e #FTE# (fim do texto) estabelecidos pelo SISA e convertidos em arquivo TXT (Apêndice C).

4.2.2 Testes

No intuito de verificar se os arquivos aplicados no SISA estavam adequadamente configurados, realizamos dois testes com a aplicação de dez artigos científicos no SISA. Foi necessário corrigir alguns detalhes relacionados à adaptação das palavras vazias e o novo teste revelou a adequação dos arquivos. No entanto, durante a aplicação do SISA verificamos que alguns termos que constam no vocabulário controlado e que contemplam os critérios de indexação automática do SISA não estavam sendo atribuídos. Em análise, constatou-se que o motivo estava relacionado ao fato de o sistema não reconhecer palavras que são constituídas por "ç" ("cê cedilhado"), uma vez que o sistema foi concebido para o idioma espanhol, que não apresenta palavras com o sinal diacrítico representado pela cedilha. Nesse sentido, foi necessário substituir o caractere "ç", que indica que a letra está cedilhada, pelo caractere "c", sem o sinal diacrítico "cedilha", em todas as palavras do vocabulário controlado, assim como nas que constituem os artigos.

4.2.3 Indexação automática dos artigos científicos

Realizamos a indexação automática dos 100 artigos científicos selecionados. Seguimos os procedimentos descritos a seguir:

1º Passo: Configuração do SISA para selecionar os arquivos de Lista de palavras vazias e Lista de descritores.

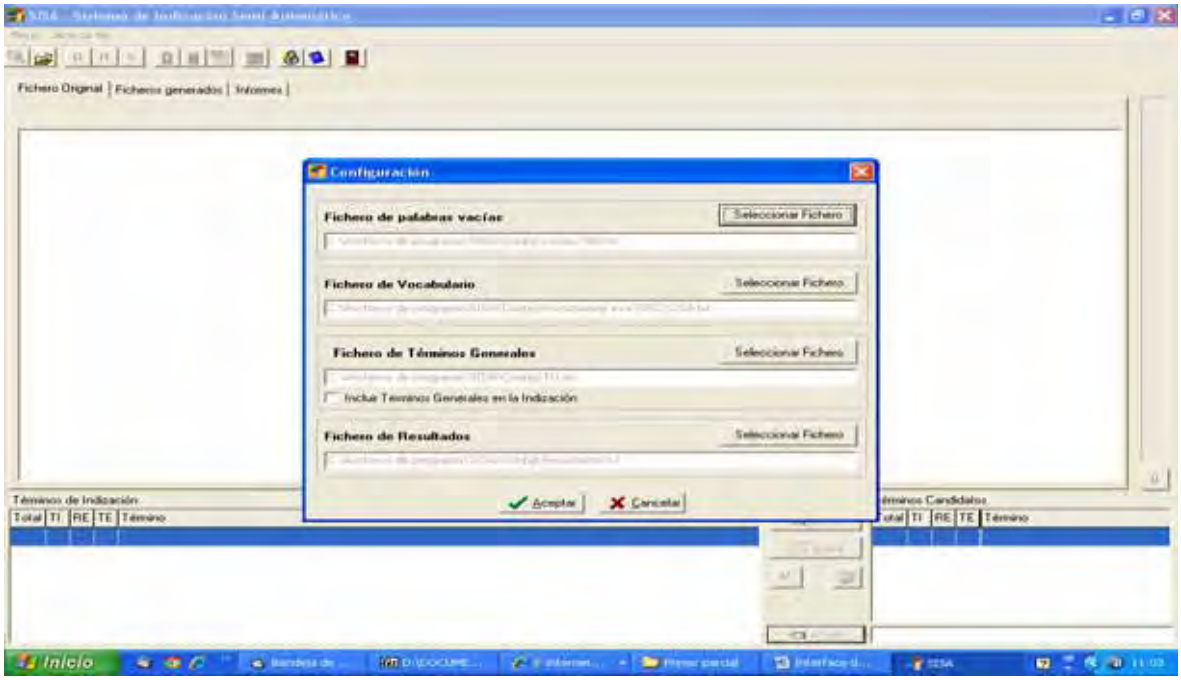


FIGURA 7 - Configuração dos arquivos no SISA

2º Passo: Configuração do SISA para selecionar dez artigos científicos no SISA.

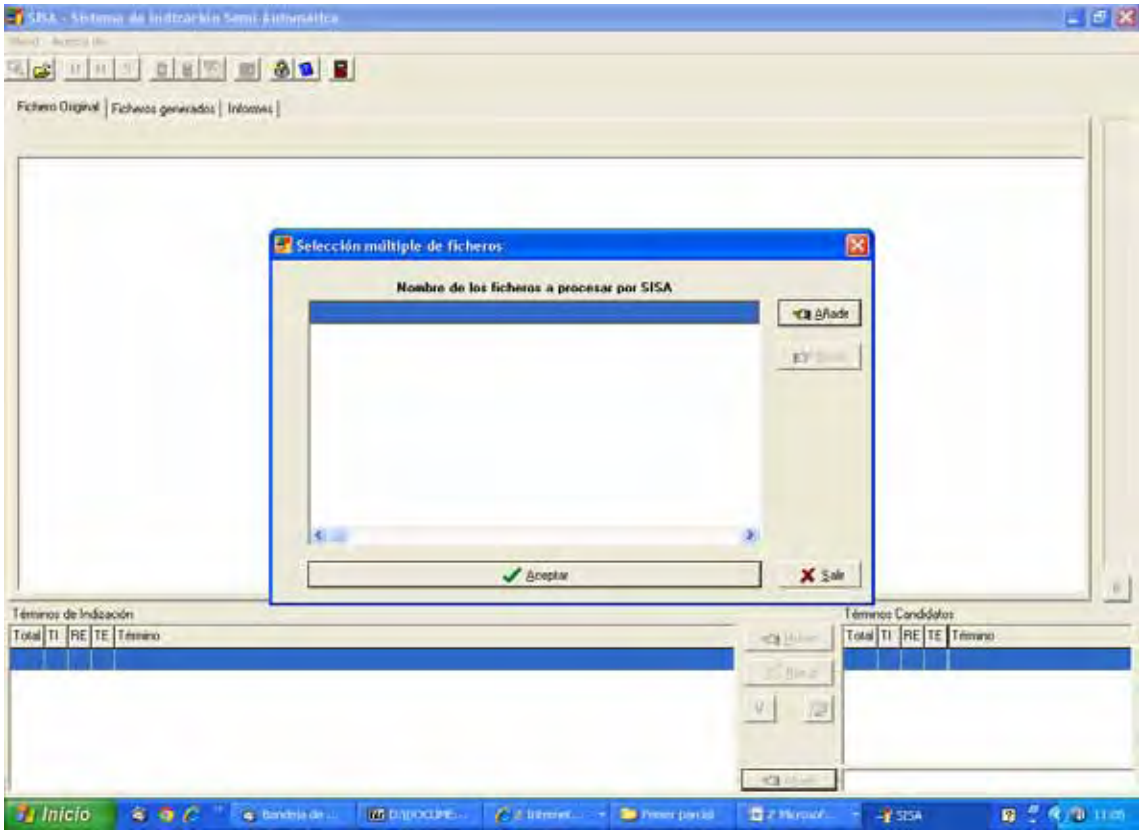


FIGURA 8 - Configuração dos artigos científicos no SISA

3º Passo: Indexar os artigos científicos no SISA

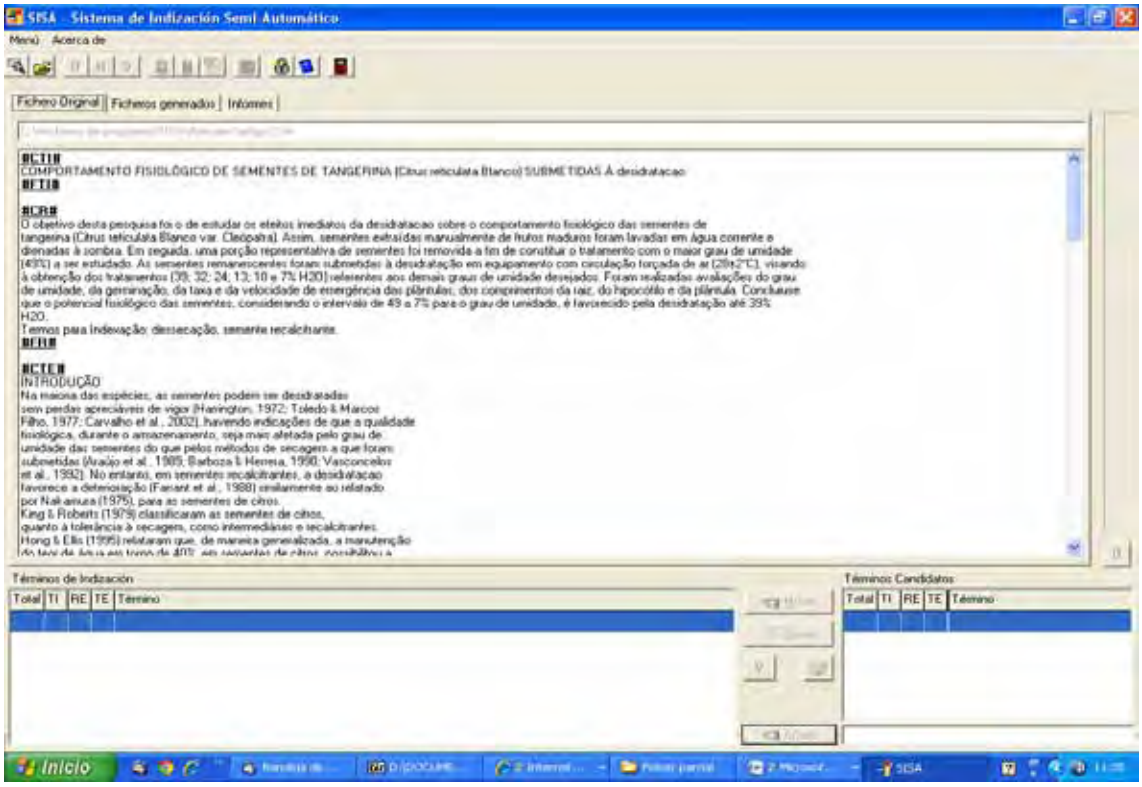


FIGURA 9 - Apresentação do sistema configurado

4º Passo: Apresentação e coleta dos termos de indexação propostos pelo SISA.

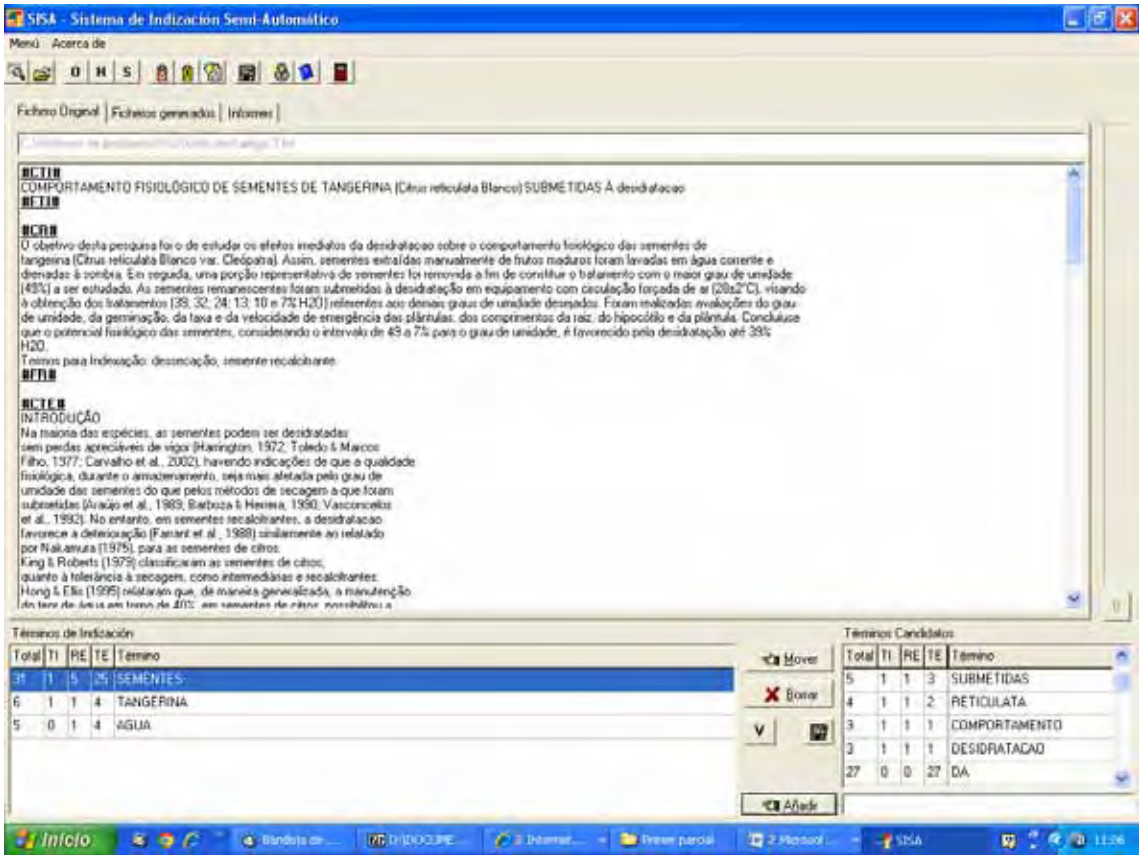


FIGURA 10- Apresentação dos termos de indexação propostos pelo SISA

5º Passo: Retorno ao 2º passo para selecionar os demais artigos científicos.

À medida em que realizamos a indexação automática de dez artigos científicos, registramos os termos de indexação obtidos por esse processo em um quadro que nos permitiu visualizar os termos atribuídos por indexação manual da BINAGRI e por indexação automática do SISA (Apêndice D). Esses dados forneceram elementos para realizar a avaliação da consistência na indexação automática e da recuperação da informação, tal como descrita em seguida.

4.3 Avaliação da indexação automática

De acordo com Abad Garcia (2005), a avaliação de uma forma geral é o processo mediante o qual tentamos obter o juízo de valor de um objeto, de uma atividade, de um processo ou de seus resultados.

A avaliação é um processo essencial para verificar em que medida os requisitos de qualidade estão sendo atendidos. A indexação, como uma atividade estratégica em qualquer

sistema de informação, requer a análise de seus resultados. Para tanto, também é necessário recorrer à avaliação de sua atividade-fim, a recuperação da informação. A avaliação do sistema de informação pode indicar se o sistema está atendendo satisfatoriamente a seus usuários e apontar os aspectos que precisam ser melhorados.

Nesse sentido, a pesquisa propõe a análise da indexação automática associando os fatores que interferiram no processo de indexação com os que afetaram a recuperação da informação, para evidenciar a exata dimensão da qualidade desse processo na recuperação da informação, sob o enfoque da atuação do vocabulário controlado.

A metodologia aplicada na avaliação do sistema SISA é baseada nos procedimentos de avaliação da indexação apresentados por Gil Leiva (1999, 2008).

Empregamos a **avaliação intrínseca** e a **avaliação extrínseca mediante recuperação**, sendo a primeira definida como o conjunto de tarefas centradas no resultado da indexação (descritores, cabeçalhos, subcabeçalhos ou identificadores) com a finalidade de conhecer a sua qualidade. A **avaliação extrínseca mediante a recuperação** objetiva comprovar o desempenho da indexação na recuperação, principalmente no que diz respeito ao atendimento das condições de exaustividade e precisão (GIL LEIVA, 2008).

A **avaliação intrínseca** pode ser qualitativa ou quantitativa. No caso desta pesquisa, aplicamos a **avaliação intrínseca quantitativa**, que se refere à reindexação de um conjunto de documentos repetindo, na medida do possível, o contexto em que se produziu a primeira indexação, para conseguir índices de consistência entre as duas formas de indexação por meio de fórmulas matemáticas (GIL LEIVA, 2008).

4.3.1 Avaliação intrínseca quantitativa: consistência na indexação

Na definição de Zunde e Dexter[20] (1969, p. 259 *apud* GIL LEIVA, 1999), a consistência da indexação é o grau de concordância na representação da informação essencial de um documento, por meio de um conjunto de termos de indexação selecionados pelos indexadores de um grupo.

Dessa forma, na aplicação do SISA, os termos de indexação atribuídos por indexação automática e os atribuídos pela BINAGRI foram organizados em um quadro comparativo para aplicar a fórmula matemática de consistência na indexação.

[20] ZUNDE, P.; DEXTER, M. E. Indexing consistency and quality. *American Documentation*, p.259-267, jul. 1969.

De acordo com Soler Monreal (2009), o índice de consistência pode ser obtido através da fórmula originalmente proposta Hooper em 1965. Com algumas variações, essa fórmula foi utilizada por Salton e McGill (1983), Lancaster (1991), Silvester Genuardi y Klingbiel (1994) e Gil Leiva (1999) para verificar a consistência entre a indexação manual e a automática. Mais recentemente, foi aplicada nas pesquisas de Soler Monreal (2009); Oliveira (2009); Narukawa, Gil Leiva e Fujita (2009); Gil Leiva, Rubi e Fujita (2008); e Gil Leiva (2001; 2002). A fórmula é apresentada da seguinte forma:

$$C_i = \frac{T_{co}}{(A+B)-T_{co}}$$

onde,
T_{co} = Número de termos comuns nas duas indexações
A = Número de termos usados na indexação A
B = Número de termos usados na indexação B

A pesquisa considera a indexação manual da BINAGRI como parâmetro de qualidade para avaliar a indexação automática do SISA, ou seja, os índices de consistência oscilarão entre 0 e 100% e, quanto mais próxima da indexação manual, maior será considerada a qualidade dos termos de indexação propostos pelo SISA.

É evidente que existem fatores distintos interferindo nos resultados de indexação manual e automática e que não podem ser negligenciados em uma análise. No entanto, o objetivo da pesquisa centra-se em avaliar a indexação resultante do processo automático, valendo-se dos resultados da indexação manual apenas como indicativo de uma indexação aceitável. A partir desse parâmetro foi possível levantar questionamentos e discussões que permitiram refletir sobre as possibilidades e limitações do emprego de vocabulário controlado no processo de indexação automática.

O QUADRO 9 mostra uma parte dos artigos selecionados com respectivos termos de indexação e o índice de consistência obtido da análise comparativa. É importante esclarecer que, para analisar a coincidência entre os termos de indexação e aplicar o valor na fórmula de consistência, considerou-se o critério de comparação “relaxada” (GIL LEIVA, 2008).

QUADRO 9 - Consistência entre a indexação elaborada pela BINAGRI e por SISA (Apêndice D).

	Artigo Científico	Descritores (Indexação BINAGRI)	Descritores (Indexação SISA)		Consistência na Indexação
1	RODRIGUES, Alexandre Couto et al. Balanço de carboidratos em gemas florais de dois genótipos de pereira sob condição de inverno ameno. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 1-4. ISSN 0100-2945.	PERA GEMA GENÓTIPO AMIDO SACAROSE CROMATOGRAFIA GASOSA CARBOIDRATO BROTAÇÃO INVERNO	INVERNO ACLIMATACAO ALTITUDE ACUCARES AMIDO ACUCAR BROTACAO CLIMA CLIMA TEMPERADO	CROMATOGRAFIA VARIEDADE FRIO FLORACAO GEMA SECA SACAROSE TRABALHO	26%
2	EINHARDT, Patrícia Milech; CORREA, Elísia Rodrigues e RASEIRA, Maria do Carmo B. Comparação entre métodos para testar a viabilidade de pólen de pessegueiro. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 5-7. ISSN 0100-2945.	PÊSSEGO CULTURA IN VITRO GERMINAÇÃO PÓLEN	POLEN PESSEGO ACUCAR ACUCAR CRISTAL AGAR AGUA ANALISE METODO ESTATISTICO PRODUCAO VEGETAL CORANTE	COR VARIEDADE FRUTIFICACAO GERMINACAO LABORATORIO MEIO DE CULTURA MICROSCOPIO METODO TRABALHO	15%
3	MARTINS, Leila e SILVA, Walter Rodrigues da. Comportamento fisiológico de sementes de tangerina (*Citrus reticulata* Blanco) submetidas à desidratação. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 8-10. ISSN 0100-2945.	TANGERINA SEMENTE DESIDRATAÇÃO UMIDADE	DESIDRATACAO TANGERINA AGUA AR EQUIPAMENTO EMERGENCIA GERMINACAO HIPOCOTILO PESQUISA	PLANTULA RAIZ SEMENTE TRATAMENTO TAXA UMIDADE VELOCIDADE	25%
	(....)				
100	COELHO, Ruimário Inácio et al. Resposta à adubação com uréia, cloreto de potássio e ácido bórico em mudas abacaxizeiro 'Smooth Cayenne'. Rev. Bras. Frutic. [online]. 2007, vol.29, n.1, pp. 161-165. ISSN 0100-2945.	ABACAXI ADUBAÇÃO FOLIAR MUDA RESPOSTA DA PLANTA MÉTODO ESTATÍSTICO	ADUBACAO ACIDO ACIDO BORICO ABACAXI POTASSIO UREIA ADUBACAO FOLIAR	ALTURA AREA FOLIAR CAULE CRESCIMENTO MASSA PLANTIO SECA	11%

Fonte: Elaborado pela autora

Isso significa que, quando há coincidência total, atribui-se o valor 1; quando a coincidência é parcial, atribui-se o valor 0,50; e quando não existe coincidência, atribui-se valor 0 (zero).

A análise dos índices de consistência conduziu a necessidade de verificar os fatores envolvidos na indexação automática. Por isso, analisamos cada um dos termos atribuídos por indexação automática e aqueles atribuídos pela BINAGRI, considerando os critérios de atribuição estabelecidos pelo SISA e verificamos as limitações e a razão de atribuição ou não de um termo de indexação. Para identificar os fatores envolvidos, foram examinados e confrontados os dados apresentados nos apêndices A e D e consultados os artigos científicos.

Como alertado por Lancaster (2004), a indexação não é uma atividade que deve ser considerada como um fim em si mesma, por isso a avaliação deve ser aplicada aos resultados dessa atividade e isso somente pode ser realizado no contexto de determinada base de dados,

em formato impresso ou eletrônico. Assim, a indexação é avaliada como bem sucedida quando permite ao usuário localizar os itens de que precisa sem ter que examinar muitos de que não precisa.

Sendo assim, após avaliar o SISA com relação à consistência, realizamos a avaliação extrínseca mediante recuperação.

4.3.2 Avaliação extrínseca: exaustividade e precisão na recuperação da informação em bases de dados

Lancaster (2004) utiliza o termo "revocação", ao invés dos termos "exaustividade na recuperação", que designa a capacidade de recuperar documentos úteis, e "precisão na recuperação", que se refere à capacidade de evitar a recuperação de documentos inúteis. Lancaster (2004) comenta que existem outras medidas de desempenho de recuperação, mas a exaustividade e a precisão na recuperação parecem ser as medidas claras que expressam resultados de qualquer busca em documentos recuperados e não recuperados.

A metodologia de avaliação extrínseca na recuperação da informação proposta nesta pesquisa oferece uma forma de comprovar em que medida a indexação automática proporcionada pelo uso do vocabulário controlado no SISA satisfaz as características de exaustividade e precisão requeridas na recuperação da informação.

A avaliação extrínseca na recuperação consiste em fazer pesquisas em duas bases de dados que contêm os mesmos campos e idênticos conteúdos, exceto os termos de indexação. A partir dos resultados da busca é possível calcular os índices de exaustividade e precisão na recuperação de informação (GIL LEIVA, 2008).

Nesse sentido, realizamos a simulação de buscas em bases de dados construídas com o resultado de indexação automática. Como em todo experimento, não é possível reproduzir as condições reais de uma busca por informação. No entanto, acreditamos que esse experimento nos oferece os indícios necessários para avaliar os resultados gerados pelo SISA.

Para tanto, foram construídas duas bases de dados: BDSISA e a BDBINAGRI. Ambas as bases foram elaboradas com uso do sistema para automação de bibliotecas *Personal Home Library* (PHL)[21], versão 8.2.

[21] É uma aplicação *Web* especialmente desenvolvida para administração de coleções e serviços de bibliotecas e centros de informações e se baseia no formato UNISIST/Unesco

Em cada base de dados elaboramos os registros dos 100 artigos científicos. Foram registrados os campos em comum e, além disso, os campos de assunto foram preenchidos com os termos de indexação de suas respectivas formas de indexação. A BDSISA foi constituída pelos registros dos artigos com os termos de indexação propostos por indexação automática do SISA e, a base BDBINAGRI, com os termos de indexação propostos por indexação realizada na BINAGRI.

Para realizar as pesquisas foi preciso estabelecer condições para controle dos resultados da busca e atribuir relevância ou não para computar os índices de exaustividade e precisão na recuperação da informação.

Dadas estas condições, estabelecemos os assuntos relevantes de cada artigo científico a partir do desenvolvimento de um processo de indexação manual, para auxiliar no estabelecimento dos artigos que são relevantes para determinada consulta nas bases de dados (QUADRO 10).

QUADRO 10 - Assuntos de cada artigo científico (Apêndice E)

Artigos científicos	Relevantes para:	Artigos científicos	Relevantes para:
Artigo 1	PÊRA; CARBOIDRATO; GEMA; CLIMA TEMPERADO; AMIDO; INVERNO	Artigo 51	PÊSSEGO; PORTA ENXERTO; CLONE
Artigo 2	PESSEGO; PÓLEN; GERMINAÇÃO	Artigo 52	AGRONEGÓCIO; MAÇÃ; RENTABILIDADE; CUSTO DE PRODUÇÃO
Artigo 3	SEMENTE; TANGERINA; DESIDRATAÇÃO	Artigo 53	POLINIZAÇÃO; LARANJA; PRODUÇÃO DE SEMENTES
Artigo 4	LICHIA; FRUTIFICAÇÃO; MATURAÇÃO; FRUTO	Artigo 54	CONDIÇÃO AMBIENTAL; LARANJA; FLORAÇÃO; INDUÇÃO; FRUTA CÍTRICA
Artigo 5	LICHIA; ANELAGEM; FLORAÇÃO; FRUTIFICAÇÃO	Artigo 55	AGROTÓXICO; MAÇÃ; CONTROLE INTEGRADO
Artigo 6	MYRTACEAE; EUGENIA INVOLUCRATA; GERMINAÇÃO; PÓLEN	Artigo 56	VINHO; UVA; VARIEDADE RESISTENTE
Artigo 7	MARACUJÁ; MATURAÇÃO; PÓS-COLHEITA; SEMENTE; GERMINAÇÃO	Artigo 57	SACAROSE; CULTURA IN VITRO; MARACUJÁ
Artigo 8	MANGABA; SEMENTE; TESTE DE VIGOR; EXTRAÇÃO	Artigo 58	MANGA; MATURAÇÃO; FRUTO; ARMAZENAMENTO; EMBALAGEM
Artigo 9	PÊRA; BORO; CÁLCIO; GEMA	Artigo 59	MARACUJÁ; PROPAGAÇÃO VEGETATIVA; ESTACA; ENRAIZAMENTO
Artigo 10	PÊSSEGO; FLORAÇÃO; BROTAÇÃO; FRUTIFICAÇÃO	Artigo 60	CARVÃO; BANANA; PROPAGAÇÃO VEGETATIVA; CULTURA IN VITRO

Fonte: Elaborado pela autora

Para determinar os assuntos relevantes, realizamos a leitura e a análise do título, do resumo, da introdução e da conclusão de cada artigo científico, com o objetivo de selecionar os seus principais assuntos e representá-los de acordo com o vocabulário controlado aplicado na pesquisa. Ou seja, realizamos um processo de indexação para que na próxima etapa fosse possível determinar para quais estratégias de busca esses artigos deveriam ser recuperados.

A segunda etapa consistiu em elaborar as necessidades de informação pesquisadas nas bases de dados. Elaboramos 50 necessidades de informação a partir da análise dos assuntos que verificamos no QUADRO10. Nessa elaboração, buscou-se contemplar pesquisas do tipo simples com descritores que representam assuntos mais amplos e com apenas um termo de indexação, bem como buscas mais complexas com dois a três termos de indexação associados, representando necessidades de informação específicas. Cada pesquisa está associada à estratégia de busca a ser executada na base de dados e aos respectivos artigos sugeridos para atender essa necessidade de informação (QUADRO 11).

É necessário esclarecer que os procedimentos para determinar os assuntos de cada artigo científico, assim como para elaborar as necessidades de informação foram pautadas por análises sistemáticas e objetivas, buscando simular um contexto com condições mínimas que se contempla em uma pesquisa em base de dados[22].

QUADRO 11 - Necessidades de informação e respectivos artigos científicos relevantes nas bases de dados (Apêndice F)

Necessidades de informação:	Estratégia de busca	Artigos relevantes nas bases de dados:
1. Artigos sobre Adubação de bananeiras	*Adubação* E *Banana*	Artigos 25; 37; 98 e 99
2. Artigos sobre Fertirrigação com potássio	*Fertirrigação* E *Potássio*	Artigos 26; 67; 98 e 99
3. Artigos sobre Adubação verde	*Adubação verde*	Artigo 15
4. Artigos sobre Análise foliar de bananeiras	*Análise foliar* E *Banana*	Artigos 25 e 68
5. Artigos sobre maturação em pós-colheita	*Maturação* E *Pós-colheita*	Artigos 7; 11; 12; 16; 70 e 72
6. Artigos sobre armazenamento de frutos em pós-colheita	*Armazenamento* E *Pós-colheita*	Artigos 11; 12; 42; 43; 70 e 88
7. Artigos sobre conservação de frutos em pós-colheita	*Preservação de alimento* E *Pós-colheita*	Artigos 12; 14; 41; 42; 43; 69 e 82
8. Artigos sobre pós-colheita de manga	*Pós-colheita* E *Manga*	Artigos 14; 82 e 88
9. Artigos sobre armazenamento de pitangas	*Armazenamento* E *Pitanga*	Artigos 11 e 12
10. Artigos sobre porta-enxerto de pêssegos	*Porta enxerto* E *Pêssego*	Artigos 51 e 64

Fonte: Elaborado pela autora

[22]O experimento foi, na medida do possível, controlado, mas está suscetível a uma margem de erro, justamente porque na indexação e na recuperação da informação estão envolvidas diversas variáveis.

Em seguida, realizamos as buscas, inicialmente na base de dados BDSISA e, depois, na BDBINAGRI. A estratégia de busca utilizada nas bases de dados foi executada segundo orientação do próprio sistema PHL, que recomenda o uso de operador booleano "AND" com a ativação da ferramenta de busca "expressão", que pesquisa todos os registros que contêm a expressão de busca fornecida. Como configuração de preferências na ferramenta de busca, definimos a pesquisa em "*Índice de assuntos*", e, como campo de dados, "*Assunto*".

Em cada busca realizada aplicamos as fórmulas para os cálculos dos índices de exaustividade e de precisão na recuperação da informação, assim expressas (GIL LEIVA, 1999, 2008; LANCASTER, 2002):

$$Exaustividade = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ de\ documentos\ relevantes\ na\ coleção}$$

$$Precisão = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ recuperados}$$

O índice de exaustividade na recuperação é obtido através da relação entre os documentos relevantes recuperados e o total de documentos relevantes que se encontra na coleção completa. O índice de precisão na recuperação se obtém da relação entre os documentos relevantes recuperados e o total de documentos recuperados.

É necessário lembrar que a análise da consistência na indexação já nos ofereceu indícios sobre os resultados na recuperação, uma vez que a qualidade da indexação reflete diretamente no momento de buscar e recuperar informação. Desse modo, a avaliação extrínseca equivale à sustentação e confirmação da análise da qualidade na indexação.

Portanto, as pesquisas e os cálculos efetuados foram sistematizados da seguinte forma:

QUADRO 12 - Cálculos de exaustividade e precisão na recupera ção de informação em base de dados BDSISA e BDBINAGRI (Apêndice G)

Base de dados BDSISA (Indexação A)	**Base de dados BDBINAGRI (Indexação B)**	**Base de dados BDSISA (Indexação A)**	**Base de dados BDBINAGRI (Indexação B)**
1ª Busca: *Adubação* E *Banana* Artigos relevantes: 25; 37; 98 e 99 Recuperados: 0 Exaustividade = 0/4 = 0% Precisão = 0/0 = 0%	1ª Busca: *Adubação* E *Banana* Artigos relevantes: 25; 37; 98 e 99 Recuperados: 37; 98 e 99 Exaustividade = 3/4 = 0,75 = 75% Precisão = 3/3= 1 = 100%	2ª Busca: *Fertirrigação* E *Potássio* Artigos relevantes: 26; 67; 98 e 99 Recuperados: 0 Exaustividade = 0/4= 0% Precisão = 0/0 = 0%	2ª Busca: *Fertirrigação* E *Potássio* Artigos relevantes: 26; 67; 98 e 99 Recuperados: 26; 98 e 99 Exaustividade = 3/4= 0,75 = 75% Precisão = 3/3 = 1 = 100%
3ª Busca: *Adubação verde* Artigos relevantes: 15 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	3ª Busca: *Adubação verde* Artigos relevantes: 15 Recuperados: 15 Exaustividade = 1/1=1= 100% Precisão = 1/1=1= 100%	4ª Busca: *Análise foliar* E *Banana* Artigos relevantes: 25 e 68 Recuperados: 68 Exaustividade = 1/2=0,50 = 50% Precisão =1/1= 1 = 100%	4ª Busca: *Análise foliar* E *Banana* Artigos relevantes: 25 e 68 Recuperados: 25; 68 e 99 Exaustividade = 2/2= 1= 100% Precisão =2/3= 0,66 = 66%
5ª Busca: *Maturação* E *Pós-colheita* Artigos relevantes: 7; 11; 12; 16; 70 e 72 Recuperados: 14; 16; 69; 70 e 72 Exaustividade = 3/6= 0,5= 50% Precisão =3/5=0,6 = 60%	5ª Busca: *Maturação* E *Pós-colheita* Artigos relevantes: 7; 11; 12; 16; 70 e 72 Recuperados: 7; 12; 14; 16 e 70 Exaustividade = 4/6= 0,66= 66% Precisão =4/5= 0,8= 80%	6ª Busca: *Armazenamento* E *Pós-colheita* Artigos relevantes: 11; 12; 42; 43; 70 e 88 Recuperados: 7; 11; 12; 14; 41; 43; 69 e 70 Exaustividade = 4/6=0,66 = 66% Precisão = 4/8= 0,5 = 50%	6ª Busca: *Armazenamento* E *Pós-colheita* Artigos relevantes: 11; 12; 42; 43; 70 e 88 Recuperados: 7; 11; 12; 14; 16; 41; 43 e 70 Exaustividade = 4/6= 0,66 = 66% Precisão = 4/8= 0,5 = 50%

Fonte: Elaborado pela autora

Assim como na análise de consistência na indexação, realizamos a análise dos fatores que poderiam justificar os valores de exaustividade e precisão alcançados no experimento. Foram analisados simultaneamente o Apêndice G, que contém os cálculos de exaustividade e precisão em cada busca, e o Apêndice D, que apresenta os termos de indexação atribuídos pelo SISA e pela BINAGRI. Eventualmente, foi necessário analisar os artigos científicos para confirmar a análise dos fatores que interferiram na recuperação da informação. Além disso, recorremos aos fatores identificados na análise de consistência por constatar que os fatores que interferiram na recuperação estão intimamente associados aos identificados na análise de consistência na indexação.

Assim, todos os fatores intervenientes identificados foram analisados a partir dos aportes teóricos apresentados e nos conduziram a sugerir alguns aspectos que merecem melhor análise no momento de adaptar vocabulários controlados para indexação automática.

5 SISTEMAS DE INDEXAÇÃO AUTOMÁTICA

As propostas de sistemas de indexação automática identificadas na literatura de Ciência da Informação e áreas afins são apresentadas buscando elucidar os aspectos que os definem em relação à aplicação de vocabulários controlados.

Para contextualizar e ampliar a compreensão sobre os sistemas de indexação automática, estes foram selecionados e apresentados por três critérios: importância histórica, proposta metodológica e uso de vocabulário controlado em indexação por atribuição, como apresentou o QUADRO 8 no capítulo anterior.

Sendo assim, apresentamos inicialmente os sistemas de indexação automática KWIC, KWOC e KWAC, PRECIS, POPSI, NEPHIS e LIPHIS, por sua importância histórica.

Em seguida, são expostos os sistemas: SMART (identificação de termos compostos); Zstation (solução de ambiguidades); Sintagmas Nominais (KURAMOTO, 2002) (identificação de sintagmas nominais); Proposta da UTC (identificação de Unidades Terminológicas Complexas); Sintagmas Nominais (SOUZA, 2005) (identificação de sintagmas nominais); SiRILiCO (análise sintática e semântica); Indexação de acórdãos (CÂMARA JÚNIOR, 2007) (indexação automática de acórdãos); Algoritmos genéticos (representação dos documentos adaptada às necessidades dos usuários); e SintagMed (indexação automática de laudos médicos), cada um pela respectiva contribuição metodológica para a indexação automática.

Por fim, são apresentados os sistemas FAIRS, AUTOMINDEX, Concept Indexer, HEPIndexer, AUTINDEX, Sistema multilíngue (POULIQUEN, STEINBERGER e IGNAT, 2003) e CADIS, que empregam vocabulários controlados por um processo de atribuição de termos de indexação. E para esquematizar as características das propostas e sistemas apresentados, expomos em um quadro síntese, a análise de suas principais características.

5.1 Sistemas de indexação automática sob a perspectiva de sua importância histórica

Aqui são apresentados os primeiros sistemas de indexação automática desenvolvidos para a elaboração de índices. Baseados na extração automática de termos da linguagem natural, esses sistemas foram utilizados na indexação semiautomática, contribuindo com o indexador humano ao auxiliar no armazenamento e na disposição dos termos para a construção de cadeias de termos nos índices.

5.1.1 KWIC, KWOC e KWAC

O primeiro sistema de indexação automática foi desenvolvido por Hans Peter Luhn, em 1953, baseado em um princípio simples: a extração de palavras significativas — conhecidas também como palavras-chave — dos títulos ou de outras partes que constituem um documento.

Esses sistemas são popularmente conhecidos como "*Key-Word in Context*" (KWIC) ou "*Key-Word In Title*" (KWIT), "*Key-Word Out of Context*" (KWOC) ou "*Key-Word Out of Title*" (KWOT) e "*Key-Word And Context*" (KWAC) ou "*Key-Word And Title*" (KWAT) e foram desenvolvidos para elaboração de índices permutados (MOREIRO GONZÁLEZ, 2004).

A criação do índice KWIC, um índice de assunto elaborado por meio das palavras-chave[23] do título dos artigos de periódicos resgata as ideias de William Frederick Poole, em 1882, com a publicação de "*Poole's Index*" (BORKO, BERNIER, 1978[24] *apud* SILVA, FUJITA, 2004). Foi também utilizado por Crestadoro como princípio norteador na construção de índices da Biblioteca Pública de Manchester (FOSKETT, 1973).

No entanto, Hans Peter Luhn foi o primeiro a aplicar esse princípio aos sistemas de indexação automática e alavancou os estudos de métodos estatísticos para indexação. Embora esses estudos fossem superficiais e pouco rigorosos, deram origem aos primeiros trabalhos de caráter morfológico e sintático de análise textual dos documentos (PINTO MOLINA, 1993). Desse modo, apresentamos em seguida as principais características desses primeiros sistemas de indexação.

O método de indexação automática para construção do índice KWIC baseia-se na extração das palavras significativas do título ou de outra parte do texto. As palavras são organizadas em ordem alfabética com posição fixa, mantendo-se, as palavras precedentes e as seguintes à palavra indexada, na mesma posição em que se encontravam. Dessa forma, a palavra extraída pode ser visualizada no índice dentro do contexto em que se encontra na estrutura do documento, o que lhe atribui um caráter contextual e garante que sejam encontradas no índice apenas associações entre termos existentes na coleção de documentos (MOREIRO GONZÁLEZ, 2004).

---

[23] O sistema de indexação por palavra-chave vem da Alemanha no final do século XVIII e norteou toda a indexação por palavra na atualidade (GUIMARÃES, 2003).

[24] BORKO, H.; BERNIER, C. *Indexing concepts and methods*. New York: Academic Press, 1978. 261 p.

O método de construção dos índices do tipo KWOC é basicamente o mesmo realizado para elaboração do índice KWIC. O que o distingue é a localização da sequência alfabética das palavras extraídas. Essas palavras extraídas são separadas das outras palavras que constituem a referida parte do documento. Neste caso, as palavras extraídas e as do titulo são isoladas, o que torna difícil a recuperação de termos compostos (MOREIRO GONZÁLEZ, 2004).

No caso do índice KWAC, ocorre o processo de extração de palavras como no KWOC. No entanto, enquanto no KWOC o lugar que a palavra extraída ocupava no título é indicado por um sinal gráfico (“*”, “...”, etc.), no índice KWAC a palavra extraída também permanece na parte considerada.

Nesse sentido, como destaca Moreiro González (2004), o critério aplicado para extração de palavras é simplesmente a sua presença na parte do documento, decisão que se fundamenta exclusivamente em elementos formais, exclui fatores semânticos e até mesmo sintáticos.

De acordo com Moreiro González (2004), os princípios nos quais a extração de palavras se sustenta estão relacionados à suposição de que os títulos são significativos e que as palavras extraídas no processo de construção dos índices são realmente válidas para representar o conteúdo de um documento. Baseia-se no princípio de que existe a possibilidade de uma palavra isolada se tornar ambígua, mas que o contexto que circunda esta palavra auxilia na definição e explicação de seu significado.

A vantagem em aplicar esse tipo de método de indexação é a facilidade em elaborar os índices. O custo é mínimo e não requer pessoal especializado, além de refletirem o conteúdo de uma coleção de documentos. Por outro lado, existem inconvenientes como: a inexistência de critério de valor entre palavras; as palavras extraídas possuem a mesma importância; não é possível considerar conceitos implícitos; não há controle de vocabulário; portanto, o método está suscetível à recuperação de documentos irrelevantes. Por esses motivos, depende de uma estabilidade terminológica da área de conhecimento, não apresenta remissivas e apresenta todos os sinônimos de um conceito como se fossem diferentes, da mesma forma que considera todas as entradas diferentes de cada uma das formas gramaticais de uma palavra (por exemplo: “gato”, “gata”, “gatos”, “gatas”). Além disso, termos compostos podem ser desestruturados, como no caso do índice KWOC e KWAC, e os títulos, estrutura a partir da qual, mais comumente, são construídos os índices, podem não refletir adequadamente os conteúdos como se acreditava (MOREIRO GONZÁLEZ, 2004).

Estudos como o de Gil Leiva e Rodríguez Muñoz (1997) mostram que dependendo da área de conhecimento os títulos não são fontes definitivas para fazer a indexação dos documentos, posto que apresentam apenas uma pequena parte dos termos representativos do conteúdo dos documentos. Devido a tais circunstâncias, é possível trabalhar apenas em uma única língua, seus resultados podem ser aproveitáveis no domínio técnico e científico dos artigos de periódicos e existe dificuldade de aplicá-los em buscas retrospectivas sobre um período mais amplo (CHAUMIER, 1986).

O método de indexação fundamentado em extração de palavras proporcionou praticidade no processo de construção de índices, principalmente com a aplicação de computadores. O indexador não tem participação intelectual, seu papel se resume a atividades operacionais, uma vez que o computador simplesmente realiza todo processo.

Os resultados oferecidos pelo método de extração de palavras mostram, sem dúvida, que sua qualidade é questionável, haja vista a gama de interferências linguísticas que se apresentam quando nos referimos ao uso da palavra como elemento de representação do conteúdo dos documentos. Segundo Silva e Fujita (2004), tais interferências começaram a ser mais intensamente discutidas, buscando investigar as interfaces da Ciência da Informação com a Linguística, justamente quando apareceram os estudos de indexação automática.

Alguns sistemas surgiram com a proposta de auxiliar o trabalho do indexador humano, oferecendo orientação quanto à construção dos índices em uma organização em cadeia, como é o caso do sistema PRECIS.

5.1.2 PREserved Context Indexing System (PRECIS)

O PRECIS é uma metodologia de indexação criada em 1968 pelo Prof. Dr. Derek Austin para a construção automática de índices de assunto da *British National Bibliography*, utilizada desde 1971 até a atualidade (FUJITA, 1989). A metodologia de indexação proposta por esse sistema lhe confere a característica de preservar o contexto do documento, mediante um conjunto de procedimentos organizados dentro de suas estruturas sintática e semântica que permite não apenas a construção automática de índices de assunto, mas também a constituição de tesauros, como proposto por Fujita (1992) na área de odontologia.

É uma metodologia de indexação amplamente aceita, aplicada a diferentes áreas do conhecimento, como Administração, Matemática, Medicina, Ciências Sociais, Música, Artes Visuais e Artes do Espetáculo; diferentes tipos de documentos, como bibliografias

multidisciplinares de âmbito nacional, livros, teses, artigos de periódicos e materiais especiais (filmes, fotografias, microfilmes, microfichas, material audiovisual); em diferentes idiomas: francês, alemão, dinamarquês, polonês e português; em redes de biblioteca para intercâmbio da informação e para catálogos de assuntos de bibliotecas e respectivos tesauros para controle de vocabulário (FUJITA, 1989).

A aplicação do PRECIS não se dá de forma totalmente automática; o indexador tem um papel fundamental no processo de análise do conteúdo dos documentos para indexação. O processo automático se concretiza apenas na armazenagem e utilização dos índices, uma vez que torna o processo de organização e rotação da cadeia de descritores muito mais ágil e fácil.

De acordo com Fujita (1989), o sistema de indexação PRECIS é constituído por um conjunto de operadores de função que caracterizam a *posição* e o *significado* dos termos de acordo com o contexto do enunciado de assunto do documento analisado. Esses operadores de função possuem uma função *sintática*, que estabelece a categoria em que o descritor se encontra, e uma função *semântica*, que estabelece as relações entre os descritores, constituindo-se numa instrução de computador que, quando utilizada, aciona automaticamente a posição, função e tipografia do termo em uma entrada de assunto. Desse modo, seu uso automático necessita, além dos operadores de função que funcionam como instruções para o computador, de mecanismos e códigos específicos que também serão integrados às instruções.

Os *procedimentos sintáticos* do PRECIS são realizados com a aplicação dos operadores de função, que atuam como gramática dos termos, ou categorias, que permitem que o indexador estabeleça uma classificação dos conceitos (STRAIOTO, GUIMARÃES, 2004).

Existem três grupos de operadores de função:

a) Os *operadores principais* (“0”, “1”, “2”, “3”) representam os componentes principais de um enunciado de assunto e respondem à questão “quem fez o quê, para quem e onde”, relacionados, assim, a elementos como “localização”, “objeto de uma ação”, “ação/efeito” e “agente de uma ação”, respectivamente. Além desses operadores principais, existem também os representados pelos números “4”, “5” e “6”, que identificam conceitos “extra-assunto”, tais como “ponto de vista”, “forma”, “amostra de população/região de estudo”, e “objetivo/forma”, respectivamente.

b) Os *operadores interpostos* são utilizados para inserir termos entre aqueles introduzidos pelos operadores principais e são subdivididos em:

Elementos dependentes

- operador (p): parte/propriedade

- operador (q): membro de um grupo quase genérico

- operador (r): agregado

Conceitos coordenados

- operador (g)

Conceitos interligados

- operador (s): definidor de função

- operador (t): associação atribuída pelo autor

c) Os *operadores de diferenças* são utilizados para introduzir partes de um termo composto (adjetivos) que limitam a conotação do foco (substantivo) sem determinar a posição de termos na cadeia e nem a função sintática. As diferenças podem ser de três tipos: que se referem diretamente ao foco; que se referem à outra diferença; e diferença referente a tempo.

É necessário observar, também, que as regras de formação de termos compostos na língua portuguesa não se limitam à formação "substantivo + adjetivo", podendo ocorrer composições como "curto-circuito" (adjetivo + substantivo), "couve-flor" (substantivo + substantivo), "guarda-roupa" (verbo + substantivo), "bota-fora" (verbo + advérbio) e "abaixo-assinado" (advérbio + adjetivo).

Considerar a composição dos termos compostos no processo de indexação automática é uma tarefa difícil para os sistemas. Austin, citado por Fujita (1992), propôs duas alternativas para o PRECIS: a primeira alternativa se refere ao uso de remissivas, na qual se transformam os adjetivos em substantivos e relacionam-se estes à sua forma adjetivada anterior. A segunda alternativa considera também a forma substantivada, para depois dispor individualmente cada elemento do termo composto e marcá-lo com o dispositivo "Leady Only" (LO).

A *parte semântica* do sistema PRECIS estabelece três classes de relações semânticas (FUJITA, 1992):

a) Relações de equivalência, que podem ser estabelecidas entre sinônimos e quase sinônimos;

b) Relações hierárquicas, estabelecendo relação genérica ou hierárquica todo/parte; e

c) Relações associativas, que podem ser relação de categorias cruzadas ou relação colateral ou de parentesco.

De acordo com Fujita (1992), o sistema de indexação PRECIS possui as seguintes características:

a) é um sistema de indexação alfabética de assuntos;

b) produz índices de assunto permutado ou rotacionados;

c) consiste em um conjunto de procedimentos de trabalho e não em uma lista pré-estabelecida de termos ou frases;

d) utiliza a linguagem natural do autor do documento para a construção do enunciado de assunto, da cadeia de termos e dos tesauros;

e) possui um esquema de operadores de função que atuam como mecanismos sintáticos na análise do enunciado de assunto;

f) os mecanismos semânticos possibilitam a construção de tesauros de modo flexível, permitindo a ampliação, eliminação e correção de termos a qualquer tempo;

g) a estrutura de entradas do índice de assunto é constituída de duas linhas e três posições (“guia”, “exposição” e “qualificador”);

h) essa estrutura deverá garantir os princípios de contexto/dependência e relacionamento “um a um” entre os termos da cadeia a fim de eliminar o problema da ambiguidade;

i) o tesauro (parte semântica) no PRECIS é construído a partir de termos estabelecidos durante a indexação e não a partir de uma lista pré-estabelecida de termos;

j) a produção de entradas no índice é realizada através de três formatos de entrada: formato padrão, formato transformação de predicado e formato invertido;

k) toda concepção e desenvolvimento do sistema PRECIS foi realizada visando a sua aplicação ao processamento por computador, mas também ao processamento manual; e

l) é um sistema multilíngue, passível de ser adaptado a qualquer língua natural.

Em avaliação do PRECIS, tanto como metodologia de indexação (FUJITA, 1989) quanto para construção de linguagens de indexação, como tesauros (FUJITA, 1992), os resultados mostram que o sistema PRECIS oferece menor esforço na recuperação da informação, apesar de exigir maior esforço na indexação. Papel preponderante possuem os operadores de função, que auxiliam na garantia de uma representação adequada, pois o sistema revela o contexto do documento de forma precisa, eliminando ambiguidades e

oferecendo a possibilidade de aplicação na indexação manual, bem como na indexação automática.

Dessa forma, constata-se que o sistema de indexação PRECIS exige uma análise do conteúdo do documento muito mais profunda pelo profissional indexador, ao mesmo tempo em que oferece a indicação do que é necessário para que essa representação seja adequadamente realizada. Haja vista a tamanha quantidade de recursos informacionais que precisam ser indexados em sistemas de informação como bibliotecas, centros de documentação, etc., a aplicação de uma metodologia como a do PRECIS para nortear o processo de indexação e sua associação ao uso de sistemas computacionais pode facilitar o processo de construção dos índices.

5.1.3 POPSI

O *POstulated-based Permuted Subject Indexing Language (*POPSI) surgiu em 1969, idealizado por A. Neelamegham e colaboradores no *Documentation Research and Training Center* de Bangalore, Índia. A exemplo de outros sistemas de indexação automática dessa época, foi utilizado para gerar índices automaticamente (FUJITA, 1989).

É um sistema inteiramente baseado em princípios classificatórios e que utiliza cabeçalhos de classificação como termos de entrada na produção dos índices. Esses cabeçalhos são seguidos de uma segunda linha de termos ligados em cadeia sob uma ordem preestabelecida, cuja padronização é derivada das categorias da classificação de dois pontos de Ranganathan (FUJITA, 1989, 2003; SILVA e FUJITA, 2004).

A sequência fundamental do POPSI é constituída por uma base, que é a categoria elementar ou disciplina, e pelo núcleo dos conceitos que se relacionam com a base (RIVIER, 1992).

5.1.4 NEPHIS e LIPHIS

Na mesma tônica, surgiram os sistemas NEPHIS e LIPHIS, ambos propostos por T. C. Craven. Craven idealizou o sistema *Nested Phrase Indexing System* (NEPHIS) em 1977. No NEPHIS o indexador humano analisa os assuntos a serem indexados e traduz os resultados dessa análise em uma sequência linear de palavras e símbolos, que é registrada em um arquivo de acesso sequencial legível por máquina. Em seguida, o NEPHIS realiza a leitura

dos arquivos registrados e utiliza cada registro para produzir um conjunto de entradas do índice permutado (CRAVEN, 1978).

O NEPHIS utiliza quatro símbolos com significados sintáticos entre os termos de um enunciado de assunto ("<", ">", "?" e "@"), os quais serão interpretados por um programa de computador a fim de produzir as permutações requeridas. Todos os termos do enunciado deverão aparecer como entrada no índice, com exceção dos termos precedidos pelo símbolo "@". O maior objetivo do sistema NEPHIS é a simplicidade (FUJITA, 1989).

A principal característica do NEPHIS é o uso de marcadores "<" e ">" para identificar a relação entre uma frase e outra, o que possibilita relacionamentos entre conceitos. Porém, a simplicidade do NEPHIS quanto à possibilidade de relacionamentos criou a necessidade de pensar em uma alternativa que contemplasse relações mais complexas, para a qual foi proposto o sistema *Linked Phrase Indexing System* (LIPHIS) (CRAVEN, 1978).

O sistema LIPHIS, tal como o NEPHIS, é um sistema de indexação de assunto permutado e foi desenvolvido para ser acessível tanto para o indexador como para o programador e, principalmente, para o usuário. Foi criado com o objetivo de manipular redes complexas de conceitos relacionados e produzir índices com maior especificidade de assuntos (CRAVEN, 1978).

No LIPHIS, o indexador insere os conceitos e o sistema prepara para cada registro uma rede. Os nós representam os conceitos e, os arcos, as relações entre conceitos. E, a partir dessa rede, o sistema gera um conjunto de entradas do índice permutado (CRAVEN,1978).

Esses tipos de sistemas oferecem ao indexador um conjunto de regras para construir cadeias com a ordenação de um número de termos interconectados a fim de formar uma frase de indexação que deverá expressar, especificamente, o assunto do documento (FUJITA, 1989).

5.2 Sistemas de indexação automática sob a perspectiva de sua proposta metodológica

Neste subitem, são apresentadas propostas metodológicas que fornecem possibilidades para solucionar alguns problemas verificados na indexação automática. Em sua maioria, estão relacionadas à identificação de aspectos linguísticos, como análise morfossintática, identificação de sintagmas nominais e solução de ambiguidades em sistemas de indexação automática.

5.2.1 Sistem for the Manipulation and Retrieval of Text (SMART)

O sistema de indexação SMART foi desenvolvido dentro das propostas de pesquisas em Recuperação da Informação do projeto SMART, de Gerald Salton, em 1961, na Universidade de Harvard, mudando em 1965 para a Universidade de Cornell.

De acordo com Lancaster (2004), o SMART foi projetado de modo a atribuir pesos numéricos aos itens, a refletir a extensão com que coincidem com os enunciados de pedidos e a apresentar esses itens ao usuário de acordo com uma ordenação por provável relevância, onde aparecem em primeiro lugar aqueles com pesos maiores. Para tanto, foram investidos anos e anos de dedicação a pesquisas para seu aperfeiçoamento, incorporando diversos critérios para indexação automática.

O processo de indexação realizado pelo SMART, de acordo com Ferneda (2003), constitui-se das seguintes etapas:

a) identificar e isolar cada palavra do texto do documento ou de sua representação (resumo, palavras-chave);

b) eliminar palavras com grande frequência e pouco valor semântico (*stopwords* ou palavras vazias), tais como preposições, artigos, etc.

c) remover afixos (prefixos e sufixos) das palavras restantes, reduzindo-as ao seu radical (processo conhecido como *stemming* ou lematização);

d) incorporar radicais (termos) aos vetores dos documentos e atribuir-lhes um peso, calculado através da medida *tf * idf*. Para se obter tal medida, procedem-se as seguintes etapas:

- Define-se a frequência de cada termo (*tf*) como sendo o número de vezes que um termo (*t*) aparece em um documento (*d*);

Como a medida *tf* apenas considera a frequência de um termo em relação a um documento e não no conjunto de documentos, existe a medida de frequência inversa do documento (*idf*), que mostra a frequência de um termo em relação ao conjunto de documentos da coleção, obtida mediante a fórmula: $idf_t = N/n_t$, onde *N* é o número de documentos e n_t o número de documentos que contêm aquele determinado termo (*t*).

- Com o valor de *tf* e de *idf* já alcançado, é possível obter o peso de um termo em relação a um documento mediante a multiplicação de *tf * idf*, como mostra a seguinte fórmula: $w_{t,d} = tf_{t,d} * idf_t$.

Podem existir termos com pesos muito abaixo da média, e esses podem ser agrupados a outros termos formando termos compostos mais específicos. A identificação dos termos compostos ocorre da seguinte forma (FERNEDA, 2003):

a) eliminam-se as palavras vazias do texto dos documentos e reduz-se cada palavra restante ao seu radical, eliminando prefixos e sufixos;

b) para cada par de radicais, verifica-se a distância entre seus componentes, que não pode ultrapassar um determinado número de palavras, sendo que pelo menos um componente de cada expressão composta deve ter uma frequência relativamente alta.

c) eliminam-se expressões compostas que possuem termos idênticos;

d) o peso de uma expressão composta é uma função dos pesos de seus componentes, e deve ser superior ao peso de cada componente tomado isoladamente.

O SMART é um sistema que possui uma proposta mais avançada, no sentido de que já incorpora elementos para uma análise mais complexa. Alia critérios estatísticos e alguns critérios de análise morfológica, como a aplicação da lematização para redução da palavra em seu radical. Há uma preocupação com a identificação de termos compostos, o que caracteriza um viés voltado à análise linguística da indexação automática, indicando que os métodos até então predominantes, os estatísticos e probabilísticos, eram insuficientes para tratar da identificação do conteúdo dos documentos.

5.2.2 Zstation: enfoque sobre fenômenos de ambiguidade

A proposta de estudo aplicando o Zstation tem enfoque sobre os problemas que são ocasionados pelo fenômeno da ambiguidade. Tal estudo é caracterizado pela preocupação de Bräscher (2002), visto que, atualmente, existe uma gama de textos completos em formato digital disponíveis na internet e as ferramentas de busca, pautadas na extração de palavras, ainda têm-se mostrado ineficientes na recuperação de informações relevantes.

Segundo Bräscher (2002), atualmente os estudos sobre tratamento e recuperação da informação baseiam-se na premissa de que as ferramentas de busca precisam considerar o conhecimento sobre o significado das expressões que são tratadas e das relações que se estabelecem entre elas, da mesma forma que devem ser capazes de tratar determinados fenômenos linguísticos, como a ambiguidade.

De acordo com Bräscher (2002), entende-se "ambiguidade como uma expressão da língua (palavra ou frase) que possui vários significados distintos, podendo, consequentemente, ser compreendida de diferentes maneiras por um receptor". A ambiguidade pode dificultar o processo de busca, uma vez que o sistema pode recuperar documentos sem relação com o que foi solicitado. Isso se reflete em um grande esforço e no tempo que será exigido do usuário.

Nesse sentido, busca-se um aperfeiçoamento dos sistemas para que reconheçam e solucionem os fenômenos de ambiguidade, processo que exige diferentes níveis de conhecimentos linguísticos e extralinguísticos (BRÄSCHER, 2002). De acordo com Bräscher (2002), a ambiguidade não é um fenômeno fácil de ser resolvido por sistemas. Para que estes reconheçam a ambiguidade é necessária a formalização da informação contextual, mas nem todo tipo de informação contextual pode ser formalizado e, portanto, nem todo tipo de ambiguidade pode ser solucionado pelos sistemas.

Segundo Bräscher (2002), as ambiguidades podem ser classificadas nos seguintes tipos, de acordo com a classificação de Fuchs:

Ambiguidade morfológica: ocasionada pela policategorização, ou seja, quando uma palavra pertence a mais de uma categoria gramatical, podendo ser um substantivo ou adjetivo ou verbo, por exemplo;

Ambiguidade lexical: ocorre quando há mais de uma interpretação possível do significado de uma unidade lexical, que pode ser provocada por homografia ou polissemia;

Ambiguidade sintática: ocorre na estruturação da frase em constituintes hierarquizados, quando se definem as ligações que se estabelecem entre os sintagmas. Por exemplo: "*Eu li a notícia sobre a greve na universidade*", que pode significar tanto que "*eu li a notícia e eu estava na universidade*" ou que "*a greve ocorre na universidade*"[25];

Ambiguidade predicativa: ocorre na interpretação das relações temáticas que articulam predicado, argumentos e participantes. Por exemplo: "*A crítica deste autor*", "autor" podendo significar tanto o objeto da crítica como o agente da crítica;

Ambiguidade semântica: ocorre quando há mais de uma interpretação possível para o relacionamento dos termos na frase. Por exemplo: "*Ela não chora mais porque ele partiu*", que pode significar que "*ela chorava porque ele havia partido*" ou que "*ela parou de chorar uma vez que ele já foi embora*";

[25] Exemplos apresentados por Bräscher (2002).

Ambiguidade pragmática: diz respeito ao cálculo dos valores enunciativos, à reconstrução desses valores, que estão ligados à situação do falante no momento da enunciação. Por exemplo: "*Paulo vai à escola*", em que não se sabe se ele é estudante ou se ele está indo à escola neste momento.

Existem vários tipos de ambiguidades e cada uma delas exige um nível de complexidade. Bräscher (2002) propôs a desambiguação mediante o tratamento sintático-semântico, utilizando gráficos conceituais como estrutura de representação do conhecimento. A utilização de gráficos conceituais se baseia na teoria dos gráficos conceituais de Sowa. Segundo essa teoria, os gráficos conceituais (GCs) constituem uma linguagem de representação do conhecimento e são formados por gráficos que possuem dois tipos de nós: os **conceitos**, "representados por retângulos ou por colchetes [CONCEITO], correspondem a conteúdos de pensamento; representam entidades, ações ou estados que possam ser descritos em termos de linguagem"; e as **relações**, "representadas por círculos com uma flecha de entrada e outra de saída ou entre parênteses => (RELAÇÃO) =>, simbolizam as ligações existentes entre os conceitos e demonstram os papéis que cada entidade desenrola" (BRÄSCHER, 2002).

Bräscher (2002) propõe a aplicação de conhecimentos sintático-semânticos organizados com base na gramática de valências de Borba para solução de ambiguidades em textos de língua portuguesa.

De acordo com Bräscher (2002), o sistema Zstation constitui-se num sistema de tratamento automático da linguagem natural que realiza a análise automática da sentença mediante a coleta de toda informação relacionada tanto à sentença quanto às propriedades semânticas e morfológicas das palavras, possíveis grupos de palavras e frases e conexões possíveis entre eles, até que o conhecimento coletado permita propor uma ou várias interpretações.

A base de conhecimento desse sistema está constituída basicamente por: *conhecimento sintático*, que são as características morfossintáticas dos elementos que representam, na estrutura superficial, uma relação predicado/argumento; função sintática desses elementos e como eles organizam-se sintaticamente; e *conhecimento semântico*, que são as características dos conceitos (traços semânticos), relações semânticas (hiperonímia, sinonímia, por exemplo) e relações temáticas (agente, ação, objeto, entre outras) (BRÄSCHER, 2002).

O sistema é constituído por vários módulos especialistas (BRASCHER, 2002):

Módulo I (Geração morfossintática): identifica a qual modelo morfológico um lema morfológico está associado e, por meio de uma gramática de geração associada ao modelo, gera as diversas formas possíveis do lema;

Módulo II (Análise morfossintática): identifica o lema morfológico para cada forma no texto e sua categoria morfossintática (substantivo, verbo, pronome, adjetivo, etc.);

Módulo III (Análise sintagmática): Extrai todos os tipos de grupos necessários para a análise sintática da sentença ou de unidades de texto maiores;

Módulo IV (Análise semântica): identifica todos os conceitos associados a um lema morfológico para obter as informações semânticas necessárias. Em um segundo momento, o módulo determina todas as restrições semânticas associadas ao conceito e os parâmetros semânticos são definidos sob a forma de traços individuais e de classes e, assim, estruturados em redes semânticas.

Para realizar esse processamento, o sistema inclui ferramentas linguísticas, como um dicionário automático, constituído de um conjunto de lemas e de dados linguísticos referentes a eles; uma gramática morfológica, que inclui o modelo morfológico, a categoria gramatical, as variáveis (pessoa e tempo para verbos e gênero e número para as demais categorias às quais se aplicam) e a regra morfológica a ser aplicada em cada entrada; a gramática de argumentos, que especifica como se efetuam as ligações entre os constituintes relacionados a determinada função sintática; e a ontologia, que especifica a relação temática definida num argumento (BRÄSCHER, 2002).

O conjunto de dados registrados no dicionário, na gramática morfológica, na gramática de argumentos e na ontologia é utilizado para efetuar-se o tratamento sintático-semântico de enunciados do corpus de pesquisa, verificando-se a ocorrência de ambiguidades e se estas foram solucionadas ou não pelo sistema Zstation. Sistemas de recuperação que adotam extração de palavras por meio de métodos estatísticos e aqueles que aplicam análise sintática para extração de sintagmas exigem menor esforço do que os sistemas que incorporam tratamento semântico. Apesar disso, não são capazes de solucionar problemas linguísticos como a ambiguidade e a sinonímia, tratadas nos sistemas tradicionais que utilizam linguagem de indexação (BRÄSCHER, 2002).

A proposta de Bräscher (2002) caracteriza-se como um fator primordial no processo de indexação automática, ainda mais quando nos referimos ao vasto ambiente *Web*, que, em comparação aos ambientes especializados, está mais suscetível à diversidade de conhecimentos e, consequentemente, diante de contextos em que podem ocorrer comumente

fenômenos de ambiguidade. Sistemas automáticos só são capazes de identificar e solucionar ambiguidades quando existem ferramentas linguísticas especialmente desenvolvidas com essa proposta.

Nesse sentido, novas alternativas estão sendo investigadas, como, por exemplo, a indexação por sintagmas nominais, desenvolvida por Kuramoto (2002, 2006).

5.2.3 Modelo de indexação automática utilizando sintagmas nominais

Kuramoto (2002, 2006) verificou que, atualmente, um dos grandes desafios para permitir acesso à informação de forma precisa, mesmo em um contexto de avanço das tecnologias da informação, são os modelos utilizados pelos sistemas de recuperação da informação, fundamentados no uso das palavras como forma de acesso à informação. A preocupação se deve ao fato de que sistemas baseados na extração de palavras realizam um processo que desconstrói o trabalho intelectual do autor do documento, pois as palavras perdem o valor que lhes foi atribuído, ou seja, tornam-se carentes de "[...] referência a um objeto ou fato da realidade extralinguística do autor" (KURAMOTO, 2006, p. 126).

Para Kuramoto (2006), os sistemas de indexação automática deveriam reconhecer e extrair unidades do discurso, ao invés de basear-se na extração de palavras. Por isso, apresenta e discute um novo modelo de tratamento e indexação da informação baseado no processamento automático da linguagem natural a partir das unidades conhecidas como "sintagmas nominais" (SN). Tais unidades são compreendidas como a menor parte do discurso portadora de informação, que, ao ser extraída do texto, mantém o seu significado (KURAMOTO, 2006).

Kuramoto (2002, 2006) propôs um modelo de reconhecimento e extração de SN baseado na estrutura sintática. Verificou que a utilização de SN na recuperação da informação pode ser realizada sob duas formas: a) implementar uma indexação automática, substituindo o modelo de indexação baseado em palavras pelo modelo baseado em SN; b) aproveitar a organização hierárquica, em árvore, dos SN para criar um novo conceito em termos de indexação, como, também, para inovar em termos de interface de busca.

O modelo foi elaborado a partir da experimentação de 15 artigos científicos da revista "Ciência da Informação", em que, sob uma abordagem lógico-semântica, foram extraídos 8.800 SN, sendo que 6.010 sintagmas de ocorrência não repetitiva foram selecionados, e

estabeleceu-se um conjunto de regras para descrever a sua estrutura sintática, definindo, dessa forma, a regra básica para descrever um SN (KURAMOTO, 2006).

Os SN são organizados em níveis hierárquicos, o que oferece uma interface de busca interessante para o usuário. Como explica Kuramoto (2006), essa interface poderia funcionar da seguinte maneira: o usuário inicialmente fornece ao sistema um termo ou palavra que representa o centro do SN que caracteriza o primeiro nível, por exemplo, “informação”, e o sistema oferece todos os SN de primeiro nível que apresentam “informação” no seu centro, inclusive o sintagma “a informação”. A partir deste sintagma, o usuário seleciona o sintagma que atende a sua necessidade de informação e o sistema apresenta, por exemplo, todos os sintagmas que têm “a informação”, tais como “o estudo da informação”, “análise da informação”, “ciência da informação”, etc. O sistema apresenta outros SN de diversos níveis até que o usuário encontre o SN que representa sua necessidade de informação. Uma interface dessa natureza intensifica a interação entre o usuário e o computador, oferecendo ao usuário a oportunidade de orientar e reorientar o computador, assim como sua busca mediante interação com a máquina (KURAMOTO, 2006).

Segundo Kuramoto (2002), os resultados obtidos com a implementação desse modelo demonstraram a viabilidade técnica de implementar-se uma interface de busca capaz de navegar em uma estrutura hierárquica em árvore de SN.

Se antes a atenção se voltava à identificação de palavras, com as pesquisas em PLN para indexação automática com uso dos SN o foco tem-se voltado à identificação do contexto em que se encontram essas palavras, isto é, os métodos de indexação automática são desenvolvidos por proposição de uma análise contextual. Nesse sentido, outros pesquisadores, como Café (2003), também se preocuparam em investigar uma metodologia de identificação de conceitos constituídos por mais de uma unidade lexical.

5.2.4 Proposta metodológica para identificação de Unidades Terminológicas Complexas – UTC

Outra proposta metodológica para indexação automática foi investigada por Café (2003), que destacou a necessidade de investigar o conhecimento linguístico, uma vez que os sistemas automáticos de tratamento e recuperação da informação apresentam melhores resultados qualitativos quando utilizam recursos da Inteligência Artificial com abordagem linguística.

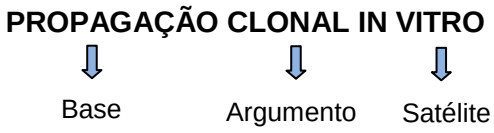
Dessa forma, à luz da interpretação de base funcionalista fundamentada na Teoria da Gramática Funcional de Simon Dik, Café (2003) propôs investigar os fenômenos linguísticos que ocorrem em unidades lexicais que representam o assunto dos documentos. Para tanto, Café (2003) utilizou um corpus formado pelo que denominou de "Unidades Terminológicas Complexas" (UTC), de base nominal do português do Brasil, a partir de textos da área de Biotecnologia de Cultura de Tecidos de Plantas.

Segundo Café (2003, p. 120), as UTC são expressões constituídas por mais de um componente lexical e sua escolha para a pesquisa se justifica pelo fato de a "maior parte da terminologia de uma área possuir uma estrutura complexa resultante da natureza onomasiológica da terminologia", o que quer dizer que, quando um especialista nomeia algo, parte de um termo-base e acrescenta a ele os elementos necessários para designar uma dada realidade (CAFÉ, 2003).

Para Café (2003), a Teoria da Gramática Funcional fundamenta sua proposta na medida em que considera a língua como um instrumento de interação social cuja função de comunicação é ressaltada e, sob a perspectiva metodológica, propõe um modelo para a análise da predicação, ou seja, a análise do processo de atribuição de predicados na expressão linguística.

De acordo com Café (2003), as UTC são definidas como unidades constituídas por uma base e por argumentos e/ou satélites. Inicialmente, Café (2003) expõe a análise semântica que mostra a função semântica de cada componente da UTC, identificando a base do segmento a partir da qual é possível determinar o papel semântico dos outros componentes da UTC, o argumento que mantém relação direta com a base e é um elemento fundamental exigido pela semântica da predicação e o elemento satélite que mantém relação com todo o conjunto de elementos à esquerda do segmento. Posteriormente, determina-se a categoria gramatical, a função sintática e a função pragmática para cada elemento da UTC.

Exemplo:



Café (2003) constatou que 47,61% dos termos complexos analisados nessa pesquisa são constituídos por dois elementos — UTC com regra de formação do tipo "base + argumento" —, o que tem sido demonstrado por outros estudos. Café (2003) também destaca

o papel semântico importante que exercem as preposições que estão ligadas ao argumento, porque a mudança da preposição pode mudar o significado da UTC, sendo que é a interpretação do papel ou função semântica de cada item lexical, à direita e à esquerda do elemento de junção, que irá definir o valor semântico-funcional da preposição.

Café (2003) ressalta que a relação entre os componentes e a sua posição na estrutura da UTC pode determinar a função semântica do item lexical e nos mostra que, entre a base e o argumento das UTC, existem relações conceituais que podem auxiliar na recuperação da informação.

Sendo assim, Café (2003) conclui que, para extrair termos realmente representativos, os sistemas automáticos precisam do domínio adequado de mecanismos linguísticos, que são bastante complexos para a análise humana e são também cruciais para a análise automática.

A identificação automática de termos formados por mais de uma unidade lexical é um fator preponderante nos sistemas de indexação automática, visto que, se não forem identificadas, a análise se fundamenta em estruturas fragmentadas que podem distorcer o real significado do conteúdo do texto, o que compromete a indexação.

5.2.5 Metodologia para atribuir descritores a partir da extração de sintagmas nominais

Souza (2005) verificou que a representação semântica dos documentos tem sido pouco explorada para automatizar e melhorar as tarefas de indexação, organização e recuperação de informação e que isso se deve à dificuldade em considerar as relações entre conceitos de forma que se considere o contexto implícito dos documentos; ou seja, fica claro que as questões linguísticas interferem no processo de indexação.

Sendo assim, Souza (2005) constata que as pesquisas nessa área podem incluir o uso de estruturas da linguagem natural, como os SN e os sintagmas verbais, e de ferramentas de representação de relacionamentos, tais como os tesauros, assim como as estratégias advindas da Linguística e da Ciência da Informação.

Nesse contexto, Souza (2005) propôs investigar o potencial dos SN em processos de indexação automática, partindo do pressuposto de que são estruturas que suportam carga semântica e podem ser utilizadas como descritores no processo de indexação automática.

A partir de pesquisas já realizadas relacionadas à utilização de SN, Souza (2005) apresentou uma metodologia para viabilizar o processo de atribuição de descritores através da

extração de SN e da análise da frequência desses sintagmas no documento e na coleção, a estrutura do sintagma, o nível dos sintagmas e a ocorrência desses em um tesauro.

De acordo com Souza (2005), a pesquisa utilizou dois *corpora*: a) 15 textos utilizados por Kuramoto (1999[26] *apud* SOUZA, 2005) em seu estudo sobre a possibilidade de aplicação de SN à indexação e recuperação da informação; e b) 60 documentos textuais de língua portuguesa selecionados de uma coleta de 75 textos que representam a totalidade dos artigos publicados em 2002 e 2003 nas revistas "DataGramaZero" e "Ciência da Informação". O procedimento realizado na pesquisa para verificar a viabilidade da indexação automática via aplicação de SN é descrito da seguinte maneira:

a) Delimitar a área em que a pesquisa seria aplicada selecionando, para tanto, os textos do campo da Ciência da Informação;
b) Os textos selecionados foram convertidos em formato de arquivo para texto simples (TXT);
c) Foram retirados os resumos e palavras-chave dos textos, atribuídos pelos autores;
d) Mediante a aplicação de ferramentas específicas para a extração de SN, foi possível obter um arquivo contendo os SN na ordem em que ocorreram nos textos originais;
e) Os SN foram ordenados de acordo com a frequência de ocorrência de cada um no corpo do documento;
f) Foram descartados os SN com frequência inferior ao estabelecido;
g) Os SN com frequência igual ou maior ao critério estabelecido são agrupados a partir de sua forma canônica e reordenados;
h) Em uma etapa opcional, é possível construir uma lista de palavras proibidas mediante a análise manual dos SN que são considerados inapropriados para indexação;
i) Verifica-se a incidência dos SN no conjunto de documentos, pressupondo-se que, quanto maior a sua incidência no conjunto de documentos, menor será sua relevância como descritor;
j) Analisar a estrutura sintática e os níveis dos SN para a análise de sua relevância;

[26] KURAMOTO, Hélio. *Proposition d'un système de recherche d'information assistée par ordinateur: avec application au portugais*. 1999. Thèse (Doctorat en Sciences de l'information et de la communication) - Université Lumière–Lyon 2, Lyon, França.

k) Verificar a ocorrência desses SN de forma parcial ou total em tesauro específico;

l) Avaliar a relevância dos SN como descritores considerando fatores como: i) a frequência de ocorrência do SN no texto do documento, ii) a incidência dos SN no conjunto de documentos, iii) seus níveis, iv) suas estruturas sintáticas e v) sua ocorrência no tesauro da área;

m) Analisar comparativamente os resumos e palavras-chave dos documentos originais e os SN atribuídos como descritores, para a avaliação da metodologia.

Para consecução da análise automática dos textos foram aplicadas algumas ferramentas. Como a nossa proposta é analisar os métodos empregados na indexação automática, nos ateremos principalmente às ferramentas que permitiram extrair os SN. Sendo assim, após a conversão dos textos para arquivo de texto simples, os textos foram submetidos ao processamento do analisador sintático (parser) denominado "Palavras", desenvolvido pela Southern University of Denmark, e também ao software "Palavras Xtractor", da Universidade do Vale do Rio dos Sinos (UNISINOS) em conjunto com a Universidade de Évora, de Portugal. Segundo Souza (2005), o processador "Palavras" utiliza um modelo gramatical chamado "Gramática de Restrições", para realizar a análise dos textos sob a perspectiva dos lexemas, dos grupos de palavras e das orações, nos níveis ortográfico, sintático e semântico. Inicialmente, essas estruturas do texto são marcadas, considerando seus aspectos morfológicos, sintáticos e semânticos, constituindo uma lista de ambiguidades por meio da qual o processador, através da aplicação sucessiva e repetida de regras, resolve as ambiguidades e classifica sintaticamente cada palavra.

A partir da aplicação do "Palavras Xtractor", o resultado do processamento dos arquivos de texto submetidos ao analisador é convertido em um conjunto de três arquivos em formato "Extensible Markup Language" (XML): um arquivo com o conjunto das palavras, que informa o número de ordem da palavra na sequência do texto; o arquivo com as categorias morfossintáticas, que informa sobre as categorias morfossintáticas de cada um dos lexemas; e o arquivo de agrupamentos, que contém informações sobre as estruturas sintáticas das sentenças do texto original, em que ocorrem os agrupamentos e em que se encontram os SN (SOUZA, 2005).

Sobre o papel que o tesauro exerce no processo de indexação automática, Souza (2005) verificou que o fato dos SN ocorrerem de forma exata no tesauro de Ciência da Informação quase nada contribuiu para lhes conferir relevância como descritores e que pouco

acrescenta o fato de ocorrerem parcialmente. Constatou diversos fatores que dificultam a aplicação do tesauro na indexação automática, entre os quais:

a) a antiguidade e a falta de atualização do tesauro utilizado;

b) a dinamicidade do campo da Ciência da Informação;

c) as características interdisciplinares das temáticas das áreas refletidas nos artigos do *corpora*, confrontadas com o foco do tesauro nas temáticas mais nucleares da Ciência da Informação;

d) a dificuldade de comparar os conceitos relacionados, através de palavras-chave ou mesmo de SN;

e) a característica geral dos tesauros de focarem conceitos amplos e genéricos — mesmo que de área especifica —, em oposição à necessidade de contextualização *ad hoc* dos descritores no escopo do texto, para o aumento de seu poder discriminatório, e de caracterização do assunto dentre as publicações de uma área; e, por fim,

f) o fato de que o tesauro, com seu conjunto de conceitos representados por palavras, difere qualitativamente do conjunto de SN, que, por possuírem semântica intrínseca, prescindem do contexto atribuído.

No caso do tesauro, o contexto de cada termo é atribuído por notas explicativas, relacionamentos ou pelo próprio fato de fazerem parte do tesauro, mas, se forem considerados isoladamente, os termos apresentam significância inferior, e, por isso, Souza (2005) decidiu abandonar o uso do tesauro como fator primordial na seleção de descritores, sugerindo seu uso como um recurso acessório para a melhoria da qualidade dos descritores selecionados.

Segundo Souza (2005), os resultados do estudo demonstraram que a metodologia possibilitou a obtenção de descritores pertinentes aos documentos, considerados eminentemente positivos, contrariando experiências anteriores declaradamente malsucedidas, baseadas na extração de estruturas sintáticas, além de considerar as dificuldades enfrentadas pela inexistência de ferramentas de extração de SN.

A proposta de Souza (2005) integra várias alternativas para o desenvolvimento da indexação automática, destacando-se os critérios de frequência de ocorrência de SN, o uso opcional de lista de palavras vazias, a verificação da ocorrência dos SN no conjunto de documentos, a realização da análise sintática e a verificação, ainda, de ocorrência parcial ou total no tesauro.

5.2.6 *Si*stema de *R*ecuperação de *I*nformação baseado em Teorias da *Li*nguística *C*omputacional e *O*ntologia – SiRILiCO

Gottschalg-Duque (2005) propôs um protótipo de indexação automática de textos eletrônicos por meio da aplicação de teorias da linguística computacional e utilização de ontologias. Tal sistema é conhecido por *Si*stema de *R*ecuperação de *I*nformação baseado em Teorias da *Li*nguística *C*omputacional e *O*ntologia (SiRILiCO) e busca o emprego de uma abordagem qualitativa pautada na extração de conteúdos semânticos, ao contrário das propostas fundamentadas em métodos estatísticos e matemáticos, que dão prioridade aos critérios de frequência de ocorrência de palavras.

Gottschalg-Duque (2005) utilizou programas como o "Palavras" (BICK[27], 1996 *apud* GOTTSCHALG-DUQUE, 2005) e o programa "Protégé" (Stanford Medical Informatics, 2005), além de desenvolver um analisador semântico chamado de "GeraOnto". O sistema SiRILiCO é constituído por três módulos: o Módulo de Processamento de Linguagem Natural (MPLN), o Módulo Gerador de Ontologias (MGO) e o Módulo Gerador de Índice (MGI).

Cada módulo possui uma função específica para o desempenho do sistema. Vejamos mais detalhadamente cada um deles.

Como explica Gottschalg-Duque (2005), o Módulo de PLN é constituído pelo analisador sintático do português, chamado "Palavras", e pelo analisador semântico desenvolvido especificamente para a pesquisa, denominado "GeraOnto", e busca analisar as frases para identificar os conceitos.

O primeiro passo para a análise é realizado pelo Sub-módulo Atomizador (SMA), que divide o texto do documento em várias partes, enviando as partes "autor", "título" e "palavras-chave" para o Sub-módulo de Ontologia Formada (SMOF) e, as frases que compõem o texto, para o Sub-módulo Sintático (SMOSi), sendo estas processadas sintaticamente. O SMOSi processa sintaticamente cada frase do texto; as estruturas com etiquetagem sintática são enviadas para o Sub-módulo Semântico (SMOSe) — que, a partir da etiquetagem sintática, procede à análise semântica, em que são identificados o núcleo proposicional — e os termos que representam agentes, objetos e instrumentos, após a etiquetagem, são enviados para o Sub-módulo de Ontologia Básica (SMOB) do Módulo Gerador de Ontologia. Tal módulo é o editor de ontologia constituído pela ontologia básica, assim como pela ontologia gerada. O SMOB é uma ontologia criada e armazenada no

[27] BICK, E. Automatic Parsing of Portuguese. In García, Laura Sánchez (ed.), *Anais* / II Encontro para o Processamento Computacional de Português Escrito e Falado. Curitiba: CEFET-PR. 1996.

Protégé, formada por um padrão referencial fundamentado na análise proposicional de Frederiksen[28] (1975 *apud* GOTTSCHALG-DUQUE, 2005), em que, com base em suas classes, é possível identificar, através das relações sintáticas entre termos, as possíveis relações semânticas. A partir do SMOF, cria-se automaticamente uma ontologia leve, mediante extração dos conceitos dos textos dos documentos, pautando-se na análise proposicional.

O Módulo Gerador de Índice é o editor de ontologias por meio do qual se verifica a ocorrência do termo da consulta em alguma proposição e identificam-se os textos que contém tal proposição. Cria-se, dessa forma, uma lista invertida de conceitos, em que, para cada conceito, há uma lista com os textos nos quais aparecem, considerando que, para cada um, pode existir mais de um termo.

De acordo com Gottschalg-Duque (2005), os modelos de Sistema de Recuperação de Informação mais utilizados (quantitativos) baseiam-se no Modelo Vetorial de Salton, verificando a compatibilidade de padrão entre as palavras-chave presentes na consulta do usuário e as presentes nos documentos que compõem a coleção indexada pelo sistema. Desse modo, Gottschalg-Duque (2005) propôs realizar um experimento de validação contrastando um Sistema de Recuperação da Informação tradicional com os resultados obtidos com o modelo SiRILiCO, mediante avaliação de precisão e revocação na recuperação da informação.

Os resultados demonstram a qualidade da indexação realizada pelo SiRILiCO, que se deu por proposições (sintagmas nominais e sintagmas verbais), identificando conceitos, e não meramente por frequência de termos. Cabe destacar, como explica Gottschalg-Duque (2005), que os modelos Vetorial e SiRILiCO não realizaram tratamento de normalização das variações linguísticas, tais como as técnica de lematização, de busca em Tesaurus ou de lista de palavras vazias. Devido às suas características de desestruturação das frases, essas técnicas podem descontextualizar um termo específico ou até mesmo dificultar o julgamento de relevância de um documento (RILOFF, 1995[29] *apud* GOTTSCHALG-DUQUE, 2005).

Sendo assim, Gottschalg-Duque (2005) conclui que o sistema SiRILiCO apresentou melhores resultados do que o modelo Vetorial, ainda que tenha apresentado alguns problemas de ruído, e, nesse sentido, mostrou-se como um modelo promissor para futuras pesquisas

[28] FREDERIKSEN, C. Representing Logical and Semantic Structure of Knowledge Aquired from Discourse. *Cognitive Psychology* 7, pp 371-458, 1975.
[29] RILOFF, E. Little words can make a big difference for text classification. *Proceedings* of the 18th annual international ACM SIGIR Conference on research and development in information retrieval, 1995.

envolvendo *web* semântica e geração de ontologias. Este sistema integra componentes mais complexos e aplica métodos de PLN com uso de ontologias.

Foram também desenvolvidas propostas de sistemas de indexação especializados, visto que áreas específicas possuem características particulares, não apenas com relação à sua terminologia, mas com relação às estruturas textuais, como na área do Direito, apresentada em seguida.

5.2.7 Sistema de Indexação Automática de Acórdãos

Câmara Júnior (2007) propôs um modelo de indexação automática de acórdãos baseado em PLN, tendo em vista a quantidade de processos judiciais nos Tribunais, buscando oferecer uma ferramenta que pudesse acelerar os trâmites processuais.

O sistema de indexação automática de acórdãos proposto por Câmara Júnior (2007) foi aplicado a acórdãos de Direito Penal. Para que fosse viável, esse sistema necessitou de um conjunto de ferramentas que se integram para formar um sistema de indexação automática de documentos de acórdãos.

Entre essas, está uma ferramenta para construção de *corpus* de língua portuguesa baseado no jargão jurídico, elaborada a partir da análise do inteiro teor dos acórdãos para aplicação em ferramentas de PLN. Essa ferramenta permite carregar um documento de acórdão em arquivo TXT, sem formatação, para ser analisado. A partir de então, o sistema realiza uma análise e a classificação morfológica, recuperando cada uma das unidades léxicas do texto, selecionadas na parte que interessa para indexação automática — “relatório”, “voto do relator” e “voto do revisor”. Em seguida, o usuário da ferramenta classifica as unidades léxicas de acordo com uma tabela que apresenta as classes morfológicas.

O próximo passo é o processo de finalização, em que, a partir da análise morfológica completa, o resultado é armazenado em memória, permitindo que o *corpus* possa ser construído incrementalmente em cada análise de acórdão selecionado.

Por fim, ocorre a consolidação do *corpus*, mediante a gravação dos resultados de cada um dos acórdãos da base, sendo convertido em um esquema utilizado pelo processador de linguagem natural do banco de dados de dicionário.

Esse dicionário é constituído pelas unidades léxicas e por suas respectivas classificações morfológicas, assim como pelas probabilidades para inferências de

ambiguidade e não ocorrência. Esse banco de dados forma o corpus em língua portuguesa utilizado pelo Qtag.

O Qtag é um analisador que realiza o PLN dos textos e, para tanto, necessita de um *corpus* da língua portuguesa que será utilizado para realizar as inferências. Ao ler um texto, esse analisador tokeniza as palavras e lhes atribui uma classe morfológica. No entanto, quando essa classe não existe, ou existe em mais de uma classe, o sistema invoca seu módulo probabilístico e é capaz de analisar quais são as outras estruturas próximas a essa palavra, tais como artigo ou adjetivo, o que pode indicar, por exemplo, que, nesse caso, a palavra é um verbo.

O Qtag foi utilizado para esse fim juntamente com uma ferramenta desenvolvida para analisar o texto e permitir a extração de diversas estruturas e sintagmas candidatas a índices dos documentos.

Essa ferramenta é constituída de duas áreas, uma para seleção dos parâmetros e outra que diz respeito à execução dos procedimentos. A primeira é formada por uma caixa de seleção com todas as classes morfológicas selecionadas para a pesquisa, assim como comandos para selecionar ou remover a seleção. Escolhendo e selecionando tal classe gramatical, é possível montar uma sequência de classes para busca no texto, da mesma forma que existem duas caixas que definem como será realizada a busca, a primeira a partir do uso do tesauro (“Utilizar Tesauro”) que determina que qualquer padrão reconhecido deverá existir no tesauro, sob pena de ser desconsiderado, havendo também uma segunda opção — “Termos Relacionados” —, que indica que, caso um termo seja selecionado, todos os termos relacionados serão exibidos nos resultados (CÂMARA JUNIOR, 2007).

Após essa configuração, o procedimento seguinte é a seleção dos acórdãos para indexação, sendo necessário que estejam em formato TXT, sem formatação, permitindo, dessa forma, que o sistema analise, aplique etiquetas (parser) no texto e busque os padrões definidos pelo sistema.

Como explica Câmara Júnior (2007), após a seleção e a análise, o procedimento seguinte é a utilização do vocabulário controlado baseado em um tesauro para a atribuição dos índices. Nesse processo ocorre o reconhecimento dos candidatos a descritores, em que, no momento da análise dos sintagmas, as unidades lexicais que são reconhecidas pelo tesauro, ou remetidas a outras estruturas do tesauro — por exemplo, os termos equivalentes, gerais, ou mais específicos —, se tornam descritores dos documentos.

Para a avaliação da viabilidade dessas ferramentas e da própria metodologia de indexação automática proposta foram utilizados os acórdãos de Direito Penal do Tribunal do Distrito Federal e Territórios, formando uma base controlada de documentos. Esses documentos foram analisados e indexados automaticamente por meio de ferramentas construídas para PLN. A avaliação da metodologia e das ferramentas utilizadas na indexação foi efetuada por meio da comparação dos resultados dessa indexação com a indexação manual realizada tradicionalmente. No contexto desse estudo, Câmara Junior (2007) verificou que a indexação automática revelou-se equivalente à indexação manual.

O sistema de indexação automática de acórdãos apresenta, assim como o SMART e o SIRILiCO, recursos mais complexos de análise automática dos documentos, aplicando ferramentas baseadas em PLN. Esse sistema atua principalmente sobre as estruturas lexicais, associado à ferramenta de análise morfológica, segundo métodos probabilísticos, permitindo que o sistema atue por meio de inferências e permitindo, ainda, a aplicação do tesauro. Daí se verifica uma preocupação maior com os aspectos linguísticos e terminológicos no processamento automático, buscando preservar a semântica dos documentos.

Em seguida, apresentamos uma proposta interessante por integrar o aperfeiçoamento da indexação dos documentos segundo os indícios que são oferecidos pelos usuários durante as buscas de informação.

5.2.8 Aplicação de algoritmos genéticos na recuperação da informação

Segundo Ferneda (2009), as limitações que existem em abordagens matemáticas para sistemas de recuperação da informação exigiram a investigação de outras abordagens. Ele propôs a aplicação de algoritmos genéticos em sistemas de recuperação da informação, na qual as possíveis representações de um mesmo documento são consideradas um tipo de "código genético" deste documento. Os algoritmos genéticos (MITCHELL, 2002[30] *apud* FERNEDA, 2009) são técnicas que simulam o processo de evolução natural em uma população de possíveis soluções para um determinado problema. A cada iteração do algoritmo (*"geração"*), um novo conjunto de estruturas é criado através da troca de informações entre estruturas selecionadas da geração anterior. O resultado tende a ser um aumento da adaptação dos indivíduos ao meio ambiente, podendo acarretar também um aumento da aptidão de toda a população a cada nova geração, aproximando-se de uma solução ótima para o problema em questão.

[30] MITCHELL, Melaine. *An introduction to genetic algorithms*. Cambridge: MIT Press, 2002. 209p.

A aplicação dos algoritmos genéticos em sistemas de informação representa uma nova forma de pensar o processo de recuperação de informação, na qual as representações dos documentos são alteradas de acordo com a necessidade de informação da comunidade de usuários, manifestada através de suas buscas (FERNEDA, 2009).

De acordo com Ferneda (2009), a aplicação de algoritmos genéticos ocorre da seguinte forma: a representação de um documento pode ser considerada o código genético em que o gene binário representado por "1" indica a presença de um termo de indexação e em que "0" indica a ausência. Por exemplo, um documento representado pelo gene binário que indica a presença dos termos "*algoritmos genéticos*", "*recuperação de informação*" e "*WEB*" e a ausência dos termos "*banco de dados*" e "*redes neurais*" como ilustra a figura a seguir:

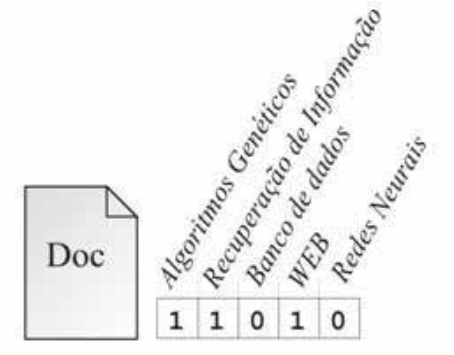


FIGURA 11 - Gene binário de um documento
Fonte: FERNEDA (2009)

A chamada "população inicial", que são os termos atribuídos a cada documento, pode ser obtida por profissionais indexadores ou mediante geração automática. O que se verifica é o caráter evolutivo nesse processo, pois ocorre uma mudança constante para que sejam progressivamente mais efetivos na identificação dos documentos.

Assim, para cada população inicial ou para cada geração nova é calculado o grau de adaptação (*fitness*) de cada indivíduo. Portanto, quando um usuário faz sua busca, esta é representada por meio de uma sequência binária, assim como os termos de indexação do documento.

O sistema então calcula o grau de adaptação (*fitness*) após a busca. Sendo assim, os indivíduos mais adaptados, ou seja, com maior *fitness*, tem maiores chances de se reproduzirem, da mesma forma que podem ocorrer processos de mutações, em que um termo de indexação existente passa a não existir em determinada posição ou vice-versa. Desse modo, todos os documentos poderão sofrer mudanças em função da expressão de busca do usuário.

De acordo com Ferneda (2009), os algoritmos genéticos representam uma nova forma de pensar o processo de tratamento e recuperação de informação, tornando-se uma alternativa ao contexto dinâmico da Web, ao permitir que as representações dos documentos se configurem adequadamente ao longo de um período, de acordo com a recuperação desses documentos por grupos de usuários com interesses comuns (FERNEDA, 2009).

A proposta dos algoritmos genéticos caracteriza-se por uma preocupação com a qualidade atribuída aos termos de indexação para recuperação de informação, uma vez que traz a concepção de que a representação deve se adaptar à relevância atribuída pelos usuários através de suas buscas. Essa concepção não é nova dentro da Ciência da Informação; no entanto, é a partir do desenvolvimento de ferramentas tecnológicas que isso pode ser concretizado com mais facilidade.

5.2.9 SintagMed

Ferneda, Galvão e Rocha (2010) apresentam a proposta de um método estatístico de indexação automática de laudos de exames radiológicos, por verificar a importância do adequado tratamento das informações na área de saúde para disponibilização rápida e precisa.

O sistema proposto foi denominado “SintagMed”. O método de indexação empregado é estatístico, com base em cálculos de similaridade, para avaliar a ligação contextual entre termos. Segundo Ferneda, Galvão e Rocha (2010), foi necessário propor uma forma de normalizar as palavras, por constatar que havia erros de digitação, excesso de abreviaturas, siglas e variações terminológicas que dificultariam a indexação automática. Dessa forma, palavras, siglas e abreviaturas foram substituídas por palavras normalizadas por meio de uma lista pré-definida contendo o termo a ser normalizado e a palavra normalizada correspondente. Também utiliza uma lista de palavras vazias.

O processo de indexação automática é realizado em duas fases. Na primeira fase, as palavras são extraídas (exceto as palavras vazias) e normalizadas. E, na segunda fase, são calculadas as forças de ligação entre as palavras, o que permite obter um conjunto de termos compostos. O sistema gera relatórios e apresenta tabelas que permitem avaliar os resultados obtidos no processo de indexação. Os resultados da análise comparativa entre a indexação manual e a automática revelaram-se promissores, por apresentar grande semelhança com relação à qualidade semântica.

5.3 Sistemas de indexação automática com uso de vocabulários controlados

As primeiras iniciativas de indexação automática estiveram fundamentadas na identificação das palavras como unidades de representação da informação. Quando a preocupação dos pesquisadores com os métodos de indexação automática volta-se aos aspectos linguísticos e terminológicos, é proposta a aplicação de instrumentos terminológicos, como tesauros agregados aos métodos de indexação automática.

5.3.1 Fully Automatic Information Storage and Retrieval System (FAIRS)

Segundo Chaumier (1986), o sistema de indexação FAIRS foi desenvolvido pelo *Centre Européen pour le Traitement de l'Information Scientifique* (CETIS) das Comunidades Europeias em Ispra. Esse sistema utiliza o método de indexação por atribuição, que consiste em realizar a indexação mediante a atribuição de termos de um tesauro previamente construído, ao contrário da indexação por extração, realizada mediante a extração dos termos do texto indexado.

Dessa forma, a indexação por atribuição depende de uma estrutura previamente definida que dê conta das inúmeras possibilidades que podem ocorrer em um texto. Chaumier (1986) demonstrou que, para considerar unitermos, assim como termos compostos, para cada termo no tesauro deverá constar o conjunto de expressões que podem aparecer associados a esse termo.

A análise automática realizada pelo sistema FAIRS é constituída pelo seguinte processo (CHAUMIER, 1986):

a) Inicialmente, analisa-se o documento frase por frase e compila-se uma lista com todas as expressões possíveis;

b) Eliminam-se dessa lista as expressões julgadas não pertinentes. Essa eliminação se dá mediante o emprego do método de ponderação, em que intervêm três fatores:

- presença de todos os unitermos de uma expressão no interior de uma mesma frase;
- peso relativo de cada unitermo dentro de uma expressão: cada expressão recebe um peso total de 100, distribuindo-se aos diferentes unitermos a componente em razão inversa da frequência de utilização dos unitermos nas expressões do tesauro;
- frequência de utilização de unitermos e expressões no texto a ser indexado;

c) Reduz-se as listas correspondentes a cada frase em uma única lista para o conjunto de texto.

Por um processo de "tradução", utilizando as relações USE e SEE ("ver") estabelecidas no tesauro, os termos da lista obtida na primeira fase são traduzidos em descritores:

Exemplo:

A USE A

A USE B

A USE A e B

A USE B e C

A USE B ou C ou D

Para que o sistema possa resolver o quinto caso, ou seja, decidir se empregará o descritor B, C ou D, aplica-se um dicionário conceitual em que os termos do tesauro são divididos em campos semânticos. Assim, a definição do campo semântico do termo será realizada novamente pelo método de ponderação: seleção dos descritores anteriormente retidos por campo semântico, acumulação dos pesos de cada descritor por campo semântico, eleição do campo semântico (para o descritor de que se trate) que tenha o peso mais elevado (CHAUMIER, 1986).

Por um lado, esse método possui a vantagem de evitar a dispersão da indexação em grande número de termos, principalmente em casos de sinonímias, e de facilitar, assim, a formulação das equações de busca (CHAUMIER, 1986). Por outro, requer a formulação de um dicionário bem estruturado e flexível o bastante para não omitir informações, oferecendo, ao mesmo tempo, certo controle para permitir que atue sob situações de ambiguidades.

Nesse sentido, identificamos também um sistema de indexação desenvolvido no Brasil, proposto por Robredo (1991).

5.3.2 AUTOMINDEX

No Brasil, o sistema de indexação "AUTOMINDEX", apresentado por Robredo (1991), possui como uma de suas características principais a existência de dois antidicionários concomitantes de palavras vazias: um de invariáveis, tais como conectivos, e outro de raízes

de palavras não significativas para determinada área do conhecimento. As vantagens de se utilizar esses dois antidicionários são: a diminuição do volume total do dicionário, a diminuição do volume de memória necessário para a armazenagem do dicionário, a diminuição do volume necessário para o processamento e o aumento da velocidade de processamento.

Robredo (1991) explica que, para o processamento, são considerados os títulos e os resumos, sendo que os caracteres a serem analisados são delimitados previamente.

Em um primeiro momento, o texto é analisado comparando-se as suas palavras com as do antidicionário de invariáveis: se forem identificadas no antidicionário, são desprezadas.

Em seguida, as palavras do texto são comparadas com o antidicionário de raízes de palavras não significativas e, as que forem identificadas, também são desprezadas.

As palavras do texto que não foram desprezadas são consideradas como possíveis descritores, ou seja, candidatas a descritores. Para definir os descritores dentre as candidatas, essas são comparadas com um dicionário de palavras significativas. Se forem identificadas nesse dicionário, são consideradas descritores, ao passo que as não identificadas permanecem sob condição de candidatas a descritores para que, em um processo de análise e avaliação pelo indexador humano, seja decidida a sua incorporação ou não como descritor. Por fim, é possível listar os descritores e candidatos a descritores com suas respectivas frequências de ocorrência na base de dados.

O AUTOMINDEX revela-se versátil para indexação de títulos e resumos de documentos, para a geração de índices temáticos e para a organização de base de dados para recuperação em linha e para a indexação de documentos correntes em arquivos empresariais, tais como cartas, ofícios, etc. (ROBREDO, 1991).

Do mesmo modo que o FAIRS, o AUTOMINDEX utiliza instrumentos que auxiliam no processo de identificar e selecionar os termos que serão utilizados como descritores; assim, tais instrumentos destinam-se a agir como filtros junto com os critérios de frequência estabelecidos no sistema de indexação.

5.3.3 Concept Assigner

Chung, Pottenger e Schatz (1998) nos apresentam a proposta de sistema de indexação automática denominado “Concept Assigner”.

O Concept Assigner é parte de um projeto maior, o “Interspace”, liderado pelo programa *Information Management*, da *Defense Advanced Research Projects Agency* (DARPA), em parceria com a CANIS, a *Community Systems Laboratory* da Universidade de Illinois, em Urbana-Champaign. A estrutura desenvolvida pretende apoiar o desenvolvimento da indexação e da recuperação da informação para os futuros repositórios.

O Interspace pretende ser um protótipo que contempla a escalabilidade, a interoperabilidade semântica interativa no domínio em questão, o tipo de mídia e o tamanho da coleção.

O Concept Assigner interage com mais dois protótipos, o “Multimedia Concept Extraction” (MCE) e o “Concept Space”, considerando que, além desses, o protótipo Interspace é constituído pelo “Category Maps” e um gerenciador geral, denominado “Domain Manager”.

O Concept Assigner executa seu papel da seguinte maneira: ao solicitar que sejam atribuídos termos de indexação a um documento, o gerenciador Domain Manager invoca o MCE para extrair os conceitos dos documentos. O MCE tem a função de extrair conceitos de vários tipos de fontes de informação multimídia, suportando fontes textuais e de imagem.

Em seguida, o gerenciador Domain Manager invoca o Concept Assigner e passa os conceitos extraídos pelo MCE. O Concept Assigner inicia, então, o processo de atribuir conceitos ao documento utilizando o Concept Space.

O Concept Space gera automaticamente vocabulários de domínios específicos que representem os conceitos e suas associações num *corpus* de informação elaborado a partir da análise de coocorrência estatística que captura a similaridade entre cada par de conceitos, partindo do pressuposto de que, quanto maior a similaridade entre os conceitos, haverá maior relevância entre eles.

O Concept Space gerado é utilizado para construir a rede Hopfield, que apresenta os nós (conceitos) e as associações/similaridade (pesos) entre conceitos.

As informações de similaridade entre conceitos obtidas a partir da análise de coocorrência estatística no Concept Space servem para o treinamento do sistema. Na medida em que um conceito extraído do documento é associado a um conceito da rede, esse conceito da rede é ativado. Assim, o Concept Space e a Rede Hopfield trabalham paralelamente, os nós da rede (conceitos) e as relações (similaridade) são atualizados sincronicamente e, quando

ocorre convergência, ou seja, quando o conjunto de conceitos tem um alto nível de ativação, são consideradas relevantes.

No experimento realizado por Chung, Pottenger e Schatz (1998), utilizou-se um conjunto de registros da base bibliográfica Compendex, produzida pela Engineering Index, que cobre a área de Engenharia.

Inicialmente, foi elaborado o Concept Space da coleção selecionada e aplicou-se o Concept Assigner para atribuir conceitos automaticamente aos documentos.

Para elaborar o Concept Space, aplicaram-se cálculos de similaridade aos campos: “Título”, “Resumo”, “Autor”, “Categorias Principais”, “Vocabulário Controlado” e “Linguagem Livre”. Os três últimos campos são índices selecionados manualmente por indexadores profissionais em Compendex.

As Categorias Principais são termos que indicam a categoria do assunto geral que se encontra no Ei Compendex Thesaurus. Para o campo “Vocabulário Controlado”, os termos são retirados de um vocabulário controlado também representado no Ei Compendex Thesaurus. Por sua vez, o campo “Linguagem Livre” é constituído por termos de linguagem livre selecionados diretamente do texto e/ou resumo do artigo.

Na avaliação realizada, os resultados do Concept Assigner indicaram menor precisão em relação ao uso de linguagem livre e vocabulário controlado; porém, apresentaram maior revocação que as demais.

Os resultados preliminares indicam o potencial existente na atribuição automática de conceitos, sugerindo que pode ser aplicado como uma alternativa para auxiliar o trabalho do indexador humano.

Esse sistema se destaca por aproveitar as relações que naturalmente se constroem nos registros de bases bibliográficas para elaborar uma rede de conceitos associados por similaridade com base na análise de coocorrência estatística. As relações entre campos de informações desses registros apresentam informações que podem evidenciar tendências em um domínio e serem aproveitadas para auxiliar na atualização de vocabulários controlados, tesauros e ontologias que serão empregados nos sistemas de indexação automática.

A identificação de relações por similaridade pode indicar relações semânticas, muitas vezes não contempladas por instrumentos de representação da informação tradicionais, e servir como um método de retroalimentação do sistema.

5.3.4 HEPindexer

O sistema HEPindexer foi desenvolvido pelo Laboratório Europeu de Física de Partículas (CERN), em Genebra, para indexar documentos sobre Física de Altas Energias. O sistema foi desenvolvido em linguagem Java, utiliza o SGBD MySQL e processa textos em formato *postscript*, PDF, ou em texto puro (ASCII), gerando lista de descritores primários e secundários (MONTEJO RAÉZ, 2001).

Assim como em outros sistemas, a motivação para seu desenvolvimento consistiu na necessidade de indexar grandes quantidades de documentos, embora esteja claro que indexadores totalmente automáticos ainda estão longe de fornecer soluções estimáveis.

Montejo Raéz (2002) identifica duas tendências, no que se refere à atribuição de descritores, e explica que o HEPindexer integra ambas:

a) Descritores para uso de indexadores humanos; e

b) Descritores propostos para uso em programas computacionais.

O HEPindexer aplica o tesauro “Deutsche Elektronen-Sybchrotron” (DESY), do laboratório alemão. O DESY traz indicações de descritores secundários, indicados por “*”; indica aqueles que não são descritores com o uso de aspas (se um termo estiver entre aspas, assim: “laboratório”, não será considerado descritor); e indica os descritores principais por meio de espaço em branco (MONTEJO RAÉZ, 2002).

Inicialmente, o algoritmo do HEPindexer realiza treinamento com um conjunto de documentos para obter dados que possibilitem o processamento posterior, isto é, o sistema trabalha com base no “aprendizado” proporcionado pela etapa anterior de treinamento com uma coleção de documentos para que esses dados sejam úteis para atribuir descritores (MONTEJO RAÉZ, 2002).

Como explica Montejo Ráez (2002), o treinamento do sistema ocorre da seguinte forma:

O sistema HEPindexer analisa o documento eliminando as palavras vazias. Em seguida, um lematizador identifica a raiz das palavras. Por fim, a frequência de ocorrência de cada termo que restou no texto é computada por meio de cálculo vetorial do termo. Após o treinamento, novo documento é alocado no sistema com o objetivo de que o sistema atribua seus descritores. O documento, então, é analisado como na fase de treinamento, atribuindo um vetor de termos por frequência. Esse vetor é multiplicado pela matriz de pesos entre descritores e termos, e oferece como resultado um vetor de descritores ponderados.

Testes mostraram uma média de 60% na precisão, assim como na revocação, e propostas de melhorias no sistema estão sendo estudadas, incluindo pesquisas sobre o uso de recursos linguísticos, inclusive recursos capazes de identificar termos compostos.

5.3.5 AUTINDEX

Segundo Nübel *et al*. (2002), o "AUTINDEX" é um sistema de indexação automática bilíngue (inglês-alemão) desenvolvido no projeto *Bilingual Automatic Parallel Indexing and Classification* (BINDEX). O objetivo do sistema é indexar automaticamente grandes quantidades de resumos de artigos técnicos e científicos da área de Engenharia.

A indexação automática utiliza um vocabulário controlado fornecido por um tesauro monolíngue e bilíngue. Também aplica a indexação automática com linguagem natural, empregada para melhorar e ampliar o tesauro. O resultado da indexação automática é uma lista de descritores e uma lista de códigos de classificação.

O sistema de indexação automática integra PLN e recursos adicionais que auxiliam no processo de representação da informação, tais como tesauro, esquema de classificação, lista de palavras vazias, lista de sinônimos e dicionário morfológico. Fornece duas alternativas de indexação: a controlada, que inclui a aplicação de um tesauro, e a livre, que se baseia apenas na análise linguística.

O analisador morfossintático opera sob três aspectos:

- Identificação da forma das palavras. Busca reconhecer limites entre sentenças, palavras, expressões e palavras constituídas por mais de uma unidade lexical para fornecer informações sobre a sequência das palavras, a categoria e a subcategoria sintática e a sequência normalizada;
- Marcação. Na marcação ou lematização as sequências são segmentadas em morfemas e esses são buscados no dicionário de morfemas, que contém informações sobre a combinação do morfema. Para evitar combinações absurdas entre morfemas, aplica-se uma lista de palavras vazias. Após o processo de segmentação, os morfemas são concatenados;
- Resolução de homógrafos. Consiste em um conjunto de regras que avaliam o contexto da palavra analisando as regularidades da palavra e a informação disponível no dicionário para reduzir a ambiguidade.

Após a análise morfossintática, ainda atua um analisador que procura resolver as ambiguidades restantes. Identifica sintagmas nominais e determina os sujeitos e os verbos no infinitivo das frases. É constituída por uma série de regras de estrutura frasal divididas em subgramáticas que são sucessivamente aplicadas.

Inicialmente, é realizada a análise do texto usando as técnicas de PLN descritas acima, para produzir uma representação sem ambiguidade e para identificar palavras compostas.

As palavras compostas são identificadas a partir de um conjunto de regras gramaticais que identificam componentes padrões e suas variações. Para calcular as palavras-chave, o sistema AUTINDEX usa função estatística de frequência. Calcula o valor que é atribuído ao substantivo em função da sua classe semântica. Isto significa que a frequência não está relacionada à simples contagem, mas é baseada na frequência das classes semânticas de cada palavra que ocorre nos documentos.

O resultado da ponderação é um conjunto de palavras-chave que pertencem às classes semânticas que foram calculadas como as classes mais frequentes.

Na próxima fase, essas palavras-chave são checadas com um tesauro fornecido por FIZ Technik (alemão) ou INSPEC (inglês). Ao atribuir os descritores do tesauro, um código de classificação está associado e pode ser usado para o propósito de eliminar a ambiguidade.

Os tesauros são formatados de modo que possam ser processados por componentes de análise linguística do AUTINDEX.

Na indexação automática bilíngue o sistema atua da seguinte forma:

- Indexa documentos em alemão com o tesauro INSPEC (tesauro em inglês), gerando descritores no idioma inglês;
- Indexa documentos em inglês com o tesauro FIZ Technik (tesauro em alemão) e gera descritores no idioma alemão.

Para gerar os termos de indexação em outro idioma são utilizados dicionários associados aos tesauros que mapeiam descritores de um idioma para outro.

5.3.6 Sistema de indexação automática de coleções multilíngues

Pouliquen, Steinberger e Ignat (2003) defendem a atribuição automática de descritores, ao invés do método de extração automática. Isso porque entendem que, em muitos casos, a extração automática de descritores inviabiliza a identificação de conceitos que não

apresentam o termo no texto. Esse método, portanto, limita, de certa forma, o processo de indexação automática.

Nesse sentido, Pouliquen, Steinberger e Ignat (2003) defendem a aplicação de tesauros conceituais, propondo o uso do tesauro Thesaurus multilingue da União Europeia (Eurovoc). O Eurovoc é um tesauro multilíngue utilizado por instituições governamentais europeias com cobertura de diversas áreas do conhecimento. A vantagem de um tesauro multilíngue é que uma coleção de documentos em diferentes idiomas pode ser recuperada por uma busca monolíngue.

Pouliquen, Steinberger e Ignat (2003) afirmam que os sistemas recentes têm trabalhado com extração de palavras e não com atribuição. No caso dos documentos que são indexados com o tesauro Eurovoc, a aplicação da extração pode ser problemática, devido à constatação de que apenas 31% dos documentos apresentavam, explicitamente, o descritor atribuído manualmente. Ou seja, os descritores do Eurovoc estão frequentemente implícitos nos documentos, e um sistema automático deve considerar essa característica.

Diante dessas implicações, Pouliquen, Steinberger e Ignat (2003) nos apresentam sua proposta de sistema de indexação automática por atribuição. Utilizaram a extração automática para obter o *corpus* de palavras-chave, necessário para que o sistema estatístico utilize o treinamento de *corpus* da indexação manual de documentos para produzir, para cada descritor, uma lista de palavras associadas da linguagem natural.

Essa proposta utilizou uma abordagem estatística com alguns recursos linguísticos de lematização, cruciais para idiomas com muitas flexões; identificação de palavras compostas, para desambiguação; e eliminação de palavras vazias, para excluir palavras insignificantes para indexação (POULIQUEN; STEINBERGER; IGNAT, 2003).

Na fase de treinamento, foi constituído um ranking de raiz de palavras (lemas) estatisticamente relacionadas a cada descritor. Para formar as listas associativas, calculou-se o valor de associação entre um lema e um descritor, por meio da fórmula de Frequência Inversa do Descritor.

Na fase de atribuição, o documento é normalizado; calcula-se então a similaridade entre a lista de frequência do lema do documento e cada lista de descritor associado. A lista de descritor associado mais similar à lista de frequência do lema do novo documento indica o descritor Eurovoc apropriado.

Os resultados por atribuição automática de descritores foram avaliados por profissionais indexadores e revelaram desempenho 570% vezes melhor em relação à extração automática de descritores. O próximo passo para a melhoria do sistema, segundo Pouliquen, Steinberger e Ignat (2003), é aplicar métodos de aprendizagem para o sistema inferir determinadas situações.

Entre os aspectos que interferem no processo de indexação automática estão também envolvidas as particularidades de cada idioma. Muitas pesquisas estão voltadas à análise de documentos no idioma inglês, e verificamos algumas que se voltam à análise de documentos em idiomas com características e complexidades diferentes, como o chinês e o árabe e o sistema apresentado em seguida, que se dedica ao idioma croata.

5.3.7 Computer Aided Document Indexing System (CADIS)

Kolar *et al*. (2005) desenvolveram o sistema de indexação conhecido como "*Computer Aided Document Indexing System*" (CADIS). A proposta do CADIS é utilizar o tesauro Eurovoc para permitir a indexação de documentos em diversos idiomas.

Esse sistema foi aplicado principalmente com a preocupação com a complexidade morfológica do idioma croata, aplicando recursos para esse idioma e para o inglês.

Os documentos no CADIS são convertidos em formato XML. Recursos linguísticos como lematização, eliminação de palavras vazias e identificação de expressões foram contemplados no sistema. Kolar *et al*. (2005) explicam a necessidade de utilizar a lematização para associar todas as palavras ao seu respectivo lema, gerando uma lista com todas as formas morfológicas encontradas no documento, com suas respectivas frequências, e uma lista contendo somente os lemas dessas palavras e a sua frequência.

Quanto à identificação de expressões, ou seja, termos compostos, por indexação automática, Kolar *et al*. (2005) empregaram o recurso N-grams, que permite identificar em uma mesma frase as palavras que se relacionam e formam termos compostos. Dessa forma, essas expressões são armazenadas no sistema e, a cada repetição nos documentos, sua frequência é acrescentada.

Kolar *et al*. (2005) explicam que o CADIS possui características multilíngues, porque o vocabulário, assim como a lista de palavras vazias, podem ser adaptados para qualquer idioma, por possuírem arquivos externos. Até mesmo a interface pode ser adaptada.

Acerca dos problemas detectados durante o desenvolvimento do sistema CADIS, Kolar *et al*. (2005) apontam que estão relacionados à complexidade morfológica do idioma croata, associada à implementação de funções estatísticas de cálculos de palavras, lemas e identificação de expressões.

Por fim, Kolar *et al*. (2005) sugerem que esse tipo de sistema pode ser utilizado como pré-processamento para atribuição de descritores, como sugestão de descritores para que o indexador humano verifique sua pertinência.

Portanto, a proposta do sistema é ser aplicado em um processo de indexação semiautomática, assim como várias das propostas apresentadas.

5.4 Análise das propostas de sistemas de indexação automática

De modo geral, podemos constatar que houve avanço significativo nas abordagens de indexação automática em que cada alternativa metodológica buscou oferecer soluções para que a qualidade na indexação pudesse ser contemplada.

Após expor alguns sistemas de indexação examinados na literatura, apresentamos um quadro que sintetiza as características verificadas durante a análise:

QUADRO 13 - Síntese das características dos sistemas de indexação automática

Ano (criação ou publicação)	**Proposta ou sistema de indexação automática**	**Características**
1953	KWIC, KWOC e KWAC	✓ Análise do título do documento ✓ Atua sobre linguagem natural ✓ Considera apenas palavras únicas ✓ Extração de palavras ✓ Facilidade na elaboração ✓ Processo totalmente automático
1961	SMART	✓ Eliminação de palavras vazias ✓ Frequência inversa do documento ✓ Identificação de termos compostos ✓ Lematização ✓ Processamento de Linguagem Natural
1968	PRECIS	✓ Construção de índices permutados ✓ Depende da indexação humana ✓ Estrutura sintática e semântica ✓ Indexação semiautomática ✓ Metodologia para elaboração de índices

(continua)

(continua)

Ano (criação ou publicação)	**Proposta ou sistema de indexação automática**	**Características**
1969	POPSI	✓ Baseia-se na categorização ✓ Construção de índices permutados
1977 **1978**	NEPHIS LIPHIS	✓ Construção de índices permutados ✓ Depende da indexação humana ✓ Estrutura sintática com marcadores de função ✓ Indexação semiautomática ✓ Manipulação de redes de conceitos
1986	FAIRS	✓ Frequência de ocorrência ✓ Indexação automática por atribuição ✓ Tesauro estruturado que associa expressões aos descritores
1991	AUTOMINDEX	✓ Análise do título e do resumo do documento ✓ Dicionário de invariáveis e dicionário de lemas insignificantes ✓ Dicionário de palavras significativas para atribuir descritores ✓ Frequência de ocorrência ✓ Indexação por atribuição
1998	Concept Assigner	✓ Indexação por atribuição ✓ Rede de conceitos formada por coocorrência de palavras ✓ Uso da rede de conceitos para atribuir descritores
2001	HEPIndexer	✓ Eliminação de palavras vazias ✓ Frequência de ocorrência ✓ Indexação por atribuição ✓ Lematização ✓ Uso de tesauro (DESY)
2002	AUTINDEX	✓ Análise morfossintática ✓ Eliminação de palavras vazias ✓ Identificação de termos compostos ✓ Indexa e classifica ✓ Indexação por atribuição ✓ Lematização ✓ Sistema monolíngue e bilíngue ✓ Uso de tesauro alemão e de tesauro inglês

(continua)

(continua)

Ano (criação ou publicação)	**Proposta ou sistema de indexação automática**	**Características**
2002	Zstation	✓ Análise morfossintática ✓ Análise sintática e semântica ✓ Dicionários de lemas ✓ Gramática de argumentos ✓ Gramática morfológica ✓ Ontologia ✓ Solução automática de ambiguidades
2002	Sintagmas Nominais (Kuramoto)	✓ Aplicação de sintagmas nominais na indexação ✓ Aplicação de sintagmas nominais na interface de busca ✓ Identificação de sintagmas nominais
2003	Proposta da UTC (Café)	✓ Análise de composição ✓ Identificação de unidades terminológicas complexas
2003	Sistema multilíngue (Pouliquen, Steinberger e Ignat)	✓ Aplicação de tesauro conceitual (Eurovoc) ✓ Eliminação de palavras vazias ✓ Frequência de ocorrência ✓ Identificação de termos compostos ✓ Indexação de conceitos implícitos ✓ Indexação por atribuição ✓ Lematização
2005	Sintagmas Nominais (Souza)	✓ Análise da frequência dos sintagmas nominais no documento e na coleção ✓ Análise sintática (ponderação da qualidade dos sintagmas nominais de acordo com sua estrutura) ✓ Identificação de sintagmas nominais ✓ Uso de tesauro
2005	CADIS	✓ Eliminação de palavras vazias ✓ Identificação de expressões ✓ Idioma croata e inglês ✓ Lematização ✓ Indexação por atribuição ✓ Indexação semiautomática ✓ Sistema multilíngue ✓ Uso do tesauro (Eurovoc)

(continua)

(conclusão)

Ano (criação ou publicação)	**Proposta ou sistema de indexação automática**	**Características**
2005	SiRILiCO	✓ Análise sintática e semântica ✓ Aplicação de ontologias ✓ Contra o uso de lematização, eliminação de palavras vazias e tesauros no sistema ✓ Extração de conteúdos semânticos ✓ Processamento de Linguagem Natural
2007	Indexação de Acórdãos (Câmara Junior)	✓ Análise morfológica ✓ Indexação de acórdãos ✓ Módulo probabilístico ✓ Processamento de Linguagem Natural ✓ Uso de tesauro
2009	Algoritmos genéticos	✓ Analogia entre a representação do documento e o código genético ✓ As buscas dos usuários como fator influenciador ✓ Representação do documento adaptada às necessidades dos usuários ✓ Representação do documento que sofre mutações
2010	SintagMed	✓ Cálculos estatísticos de similaridade ✓ Eliminação de palavras vazias ✓ Indexação de laudos médicos

Fonte: Elaborado pela autora

Os primeiros métodos de indexação automática foram desenvolvidos para a elaboração de índices que seriam impressos. Fundamentaram-se na extração de palavras do texto, utilizando a linguagem natural e a sua frequência de ocorrência para atribuir relevância, considerando estruturas específicas do documento, como o título e o resumo.

No entanto, algumas limitações que seu uso ocasiona tornaram necessária a integração de alternativas. Nesse sentido, surgem os estudos linguísticos, associando as análises morfológicas, sintáticas e semânticas e incorporando também instrumentos como listas de palavras vazias (*stopword*), dicionários morfológicos e tesauros.

Podemos deduzir que, a partir daí, surge também a concepção das bases de conhecimentos, ou seja, pressupõe-se que os sistemas precisam ter acumulado um conhecimento prévio para poder realizar a análise dos documentos. Essa concepção é

constatada nos sistemas HEPIndexer, Concept Assigner e no sistema multilíngue proposto por Pouliquen, Steinberger e Ignat (2003), que realizam um treinamento prévio com um *corpus* de documentos ou registros bibliográficos para coletar um conjunto de informações empregado para oferecer suporte na análise dos documentos.

Entre as décadas de 1960 e 1970 surgiram os sistemas de elaboração de índices permutados, como PRECIS, POPSI, NEPHIS e LIPHIS, com a proposta de indexação semiautomática, o que significa que o indexador tem papel fundamental na análise do conteúdo do documento. O sistema apenas apoia o armazenamento e a elaboração das cadeias que constituem o índice. O PRECIS, mais do que apoiar, fornece uma metodologia sintática e semântica para a construção de uma ordem lógica entre os conceitos, privilegiando o aspecto contextual. O POPSI apresenta uma abordagem sobre categorização baseada nos postulados de Ranganathan. Por sua vez, os sistemas NEPHIS e LIPHIS caminham no sentido de considerar as relações conceituais na construção dos índices.

Os sistemas de indexação ulteriores fundamentam-se em princípios mais complexos, dando atenção às questões da contextualização dos termos para representação e recuperação da informação. Essa preocupação com o contexto de significado, ou seja, com a semântica, pode ser constatada nas propostas de Kuramoto (2002), Souza (2005) e Gottschalg-Duque (2005), envolvendo a identificação automática de sintagmas nominais, e na proposta metodológica de Café (2003), ao propor o estudo das UTCs, sobretudo para a identificação dos termos constituídos por mais de uma unidade lexical para manter o real significado dos termos, aspecto extremamente relevante para a análise automática de textos. A identificação de termos compostos já vem sendo aperfeiçoada em sistemas como o SMART, o AUTINDEX, o sistema multilíngue de Pouliquen, Steinberger e Ignat (2003) e o CADIS.

Como destacado por Café (2003), na literatura técnica e científica a característica de composição é muito comum, incluída a questão da variação dos componentes dessa composição que podem ser constituídos, por exemplo, por preposição.

Desse modo, verifica-se a necessidade de uma análise sobre as características particulares do domínio em que o sistema será aplicado. Foram apresentados sistemas como o AUTINDEX, o sistema proposto por Pouliquen, Steinberger e Ignat (2003) e o CADIS, desenvolvidos especialmente para considerar as especificidades de idiomas diferentes. Além disso, o desenvolvimento de tais sistemas, assim como relatos de outros, em idiomas como chinês e árabe, são o indício da busca de uma interoperabilidade entre sistemas de diversos

países e a tentativa de amenizar as barreiras linguísticas que ainda existem quando se pensa em acesso à informação.

Quando nos referimos ao domínio em que o sistema é aplicado, podemos relacionar também as fontes de informação (documentos) geradas pelas comunidades específicas. Existem fontes de informação em diferentes formatos e padrões que precisam ser considerados, porque sua composição assinala como as informações estão apresentadas e indicam ao sistema como irá atuar para analisá-las. Entre os sistemas que analisamos, podem ser citados o de Câmara Júnior (2007), que propôs a análise de acórdãos, um tipo de documento jurídico, e o sistema SintagMed, que apoia a análise de laudos médicos.

Outra questão que envolve o domínio do conhecimento a ser tratado se refere aos instrumentos de representação da informação, tais como os vocabulários controlados, que apresentam a terminologia de um domínio. Verificamos a aplicação de vocabulários nos sistemas FAIRS, AUTOMINDEX, Concept Assigner, AUTINDEX, no sistema multilíngue proposto por Pouliquen, Steinberger e Ignat(2003), na proposta de Souza (2005), no CADIS e no sistema proposto por Câmara Júnior (2007), além do emprego de ontologias nos sistemas Zstation e SiRILiCO.

É necessário destacar que o emprego de vocabulários controlados torna-se complicado para o sistema, porque esse instrumento deve atuar com o processamento da linguagem natural. Isso significa que o sistema processa automaticamente um texto em linguagem natural, trata-o a partir de análises linguísticas para identificar os conceitos relevantes para indexação e, além disso, traduz esses conceitos em um vocabulário controlado. Nesse processo, o sistema se depara com o desafio de representar adequadamente o que foi identificado no processo de análise do conteúdo do documento.

Souza (2005) verifica que a aplicação de tesauros na indexação automática proporcionou alguns inconvenientes durante o processo de análise, devido às características pouco flexíveis desse recurso, além de problemas relacionados à atualização, limitações devidas à interdisciplinaridade do conteúdo dos documentos e às características dos conceitos.

Uma alternativa que tem sido apresentada pelos sistemas de indexação automática é integrar um instrumento de representação em uma rede de conceitos que associa um descritor a diversas expressões ou palavras. Essas relações podem ser estabelecidas, por exemplo, por meio de análise de coocorrência de termos no documento e/ou na coleção, servindo de apoio ao processo de atualização do instrumento de representação integrado ao sistema, como no sistema Concept Assigner.

Outros sistemas, como o SiRILiCO e o Zstation, aplicam ontologia, instrumento de representação de informação desenvolvido especialmente para formalizar conceitualizações, de modo que possam ser compreendidas tanto por sistemas como por humanos.

De forma geral, os atuais sistemas de indexação automática atuam por meio de análise morfológica e sintática e integram outros recursos linguísticos, como dicionários, listas de palavras vazias, tesauros ou ontologias, etc.

Entre os recursos empregados na análise morfológica é possível descrever a lista de palavras vazias e o processo de lematização, aplicados na maior parte dos sistemas analisados: SMART, HEPIndexer, AUTINDEX, sistema multilíngue proposto por Pouliquen, Steinberger e Ignat (2003), CADIS e SintagMed.

No entanto, verificamos que esses recursos, assim como a *tokenização* e o uso de tesauro, também proporcionam limitações à indexação automática. Por isso, encontramos algumas divergências de opiniões. Gottschalg-Duque (2005) destaca essa questão ao afirmar que sua proposta, o sistema SiRiLiCO, não utiliza recurso de lematização, lista de palavras vazias e tesauro, uma vez que podem desconstruir discursos durante a análise automática.

Além disso, verificamos iniciativas importantes, como a apresentada por Bräscher (2003), que se dedicou ao estudo de fenômenos linguísticos de ambiguidade, investigação que envolve a compreensão do contexto dos termos para que os sistemas automáticos possam identificá-los.

Ademais, verificamos uma concepção voltada ao usuário, com a proposta de aplicação de SN na interface de busca, sugerida por Kuramoto (2002), e com a proposta de Ferneda (2009) de aperfeiçoar a representação da informação por meio de recursos como os algoritmos genéticos, que atuam sobre os indícios obtidos nas buscas realizadas pelos usuários num processo de constante adaptação da representação da informação às necessidades dos usuários.

É no sentido de permitir que sistemas computacionais reconheçam os relacionamentos entre conceitos que os instrumentos construídos para controlar o vocabulário são requeridos e amplamente investigados. A exigência se define não apenas na construção de ferramentas computacionais, mas também na melhor forma de constituir a representação da informação. Os instrumentos de representação da informação que são tradicionalmente utilizados precisam ser adaptados a um novo contexto, com uma atuação mais flexível, e, ao mesmo tempo, requer-se que atuem como instrumento de controle vocabular. Dessa forma, os relacionamentos semânticos oferecem uma importante função na busca da garantia do

contexto informacional e podem proporcionar precisão na indexação, bem como na recuperação da informação.

Portanto, existem fatores relacionados à composição dos vocabulários controlados, tais como terminologia, estrutura e apresentação, normas de elaboração, atualização e processos de uso, que podem interferir na qualidade do processo de indexação por estarem em desacordo com os métodos de indexação automática.

Com o objetivo de aplicar o vocabulário controlado ThesAgro no processo de indexação automática do SISA, em análise comparativa com a indexação manual realizada pela BINAGRI, propomos, no próximo capítulo, investigar os resultados de indexação decorrentes dessa aplicação.

6 VOCABULÁRIO CONTROLADO NA INDEXAÇÃO AUTOMÁTICA DO SISA

Neste capítulo apresentamos os resultados do experimento em que empregamos o vocabulário ThesAgro no sistema SISA em análise comparativa com a indexação manual da BINAGRI e realizamos a simulação de buscas em bases de dados. Expomos os dados quantitativos que abrangem os índices de consistência na indexação e os índices de exaustividade e precisão na recuperação da informação e os fatores intervenientes envolvidos na indexação automática que justificam esses valores, apontando os exemplos identificados e as implicações para indexação e recuperação de informação.

6.1 Índices de consistência na indexação

Verificamos que a quantidade de descritores atribuídos em cada tipo de indexação é bastante discrepante, o que foi confirmado pelos dados: na indexação pela BINAGRI, houve uma variação de 3 a 14 descritores, com média de 6 a 7 descritores, ao passo que, na indexação automática do SISA, observamos uma variação de 4 a 25 termos atribuídos e uma média de 14 termos de indexação. Tal fato revela a influência também nos índices de consistência, já que o número de descritores atribuídos em cada tipo de indexação é considerado nos cálculos de consistência na indexação.

Verificamos uma variação de 2 a 54% de consistência na indexação dos artigos científicos, com média de 19,30% (TAB. 1). O resultado da média aritmética dos índices de consistência resultou em um valor relativamente baixo, mas é justificado a partir de análise. É necessário esclarecer que a consistência na indexação não deve ser considerada o único parâmetro de avaliação, visto que existem vários fatores que influenciam o processo de indexação.

Como observa Gil Leiva (2008), esses fatores estão relacionados ao indexador, visto que ambos, automático ou humano, possuem alguns critérios estabelecidos; ao contexto da indexação, ou seja, a política e os objetivos da indexação; à complexidade e às características do objeto a ser indexado; além de relacionados ao momento, isto é, realiza-se a comparação entre palavras-chave retiradas diretamente do texto, ou entre aquelas que foram convertidas por um vocabulário controlado.

Como salienta Gil Leiva (1999), até mesmo a exploração dos índices de consistência como indicador de uma indexação correta pode ser problemático, e isso se deve ao fato de que

podemos encontrar uma indexação consistente, porém incorreta, o que pode ocorrer por erros de análises — os dois indexadores cometem o mesmo erro —, ou pela utilização de vocabulários controlados diferentes.

Nesse sentido, as pesquisas de análise de consistência revelam a dificuldade em alcançar altos índices de consistência (LANCASTER, 2004) e apontam que, mais do que uma anormalidade, é necessário contemplar a inconsistência na indexação como um elemento inerente a esta tarefa (GIL LEIVA, 1999).

TABELA 1 - Índices de consistência na indexação

CONSISTÊNCIA NA INDEXAÇÃO							
Artigo	**Índice de consistência (%)**	**Artigo**	**Índice de consistência (%)**	**Artigo**	**Índice de consistência (%)**	**Artigo**	**Índice de consistência (%)**
1	*26*	26	*25*	51	*21*	76	*11*
2	*15*	27	*9*	52	*27*	77	*18*
3	*25*	28	*23*	53	*13*	78	*24*
4	*23*	29	*17*	54	*18*	79	*16*
5	*13*	30	*18*	55	*5*	80	*24*
6	*31*	31	*29*	56	*11*	81	*18*
7	*35*	32	*15*	57	*26*	82	*26*
8	*28*	33	*50*	58	*20*	83	*23*
9	*13*	34	*9*	59	*13*	84	*15*
10	*28*	35	*10*	60	*20*	85	*16*
11	*35*	36	*20*	61	*20*	86	*19*
12	*28*	37	*22*	62	*13*	87	*20*
13	*14*	38	*14*	63	*2*	88	*11*
14	*44*	39	*16*	64	*21*	89	*12*
15	*18*	40	*8*	65	*14*	90	*15*
16	*17*	41	*14*	66	*12*	91	*4*
17	*33*	42	*18*	67	*27*	92	*36*
18	*12*	43	*31*	68	*16*	93	*7*
19	*4*	44	*8*	69	*29*	94	*21*
20	*12*	45	*13*	70	*54*	95	*14*
21	*29*	46	*25*	71	*17*	96	*0*
22	*17*	47	*17*	72	*53*	97	*14*
23	*11*	48	*20*	73	*9*	98	*24*
24	*25*	49	*9*	74	*16*	99	*18*
25	*20*	50	*13*	75	*27*	100	*11*
Índice Médio de Consistência na Indexação						19, 30 x 100%= 1930% 1930% /100 artigos = 19,30%	

Vários estudos revelam que a margem de consistência obtida nos resultados de investigações está entre, aproximadamente, 10% e 60% (LANCASTER, 1968; LEONARD, 1975; FUNK e REID, 1983; MARKEY, 1984; MIDDLETON, 1984; TONTA, 1991; SIEVER e ANDREWS, 1991; IIVONEN e KIVIMÄKI, 1998; LEININGER, 2000; GIL LEIVA, 2001 e 2002; SAARTI, 2002; NESHAT e HORRI, 2006; GIL LEIVA, POLSINELLI e SPOTTI, 2008 e KIPP, 2009).

Considerando a margem de consistência verificada nos estudos, o índice médio de 19,30% de consistência obtida nesta pesquisa está dentro dos padrões verificados.O dado em si não revela as causas que geraram esse resultado, por isso em análise, constatamos que os seguintes fatores podem explicar melhor as interferências na indexação automática.

6.1.1 Fatores intervenientes nos índices de consistência na indexação

De modo geral, o sistema SISA não atribuiu muitos termos de indexação importantes. Da análise dos termos atribuídos pelo SISA constatamos situações frequentes que impossibilitaram a atribuição de termos de indexação que foram propostos na indexação realizada por análise humana. Além disso, ocorreram situações em que o SISA atribuiu termos dispensáveis para a indexação, o que, consequentemente, interferiu no número de termos atribuídos considerado no cálculo de consistência na indexação. Sob a perspectiva da análise semântica, verificamos também inconsistência entre o significado dos conceitos do vocabulário controlado e dos termos correspondentes analisados automaticamente no texto dos artigos científicos e atribuídos pelo SISA.

Nesse sentido, apresentamos as circunstâncias que impossibilitaram a atribuição de termos de indexação pelo sistema SISA, esclarecendo que, para compreendê-las, é necessário considerar que o sistema apenas atribui termos do artigo científico que constam no vocabulário controlado. Por um processo de comparação de padrões identificam-se os termos de indexação autorizados, considerando sua presença em combinação nas partes “título”, “resumo” e “texto”, que constituem os documentos. Foram identificados os fatores que são detalhados em seguida:[31]

- ✓ Termos no singular e no plural;
- ✓ Frequência de ocorrência dos termos em apenas uma estrutura do documento;

[31] Os termos dos artigos e do vocabulário controlado citados nos quadros não foram atualizados para a nova ortografia: os termos dos artigos foram citados tal como encontrados nos originais e os termos do vocabulário controlado são padronizados sem acentuação.

- ✓ Dificuldade em atribuir termos compostos;
- ✓ Diferença na apresentação entre os termos do artigo e do vocabulário controlado;
- ✓ Dificuldade em atribuir conceitos implícitos;
- ✓ Diferença semântica nos termos de indexação;
- ✓ Atribuição automática de termo geral e de termo específico;
- ✓ Atribuição de termos relacionados à metodologia da pesquisa;
- ✓ Relação de equivalência omitida.

A variação entre o singular e o plural de um mesmo conceito foi um dos fatores que ocorreram com mais frequência.

QUADRO 14 - Fator de interferência na indexação automática (flexão de número nos termos de indexação)

Fator	**Exemplos**	
Termos no singular e no plural	**Artigo científico**	**Vocabulário controlado**
	Agrotóxicos	Agrotoxico
	Biofilmes (resumo) e Biofilme (texto)	Biofilme
	Carboidratos	Carboidrato
	Carotenóides	Carotenoide
	Clones	Clone
	Compostos fenólicos	Composto fenolico
	Cromossomos	Cromossomo
	Custos de Produção	Custo de Producao
	Fitorreguladores	Fitorregulador USE Regulador de crescimento
	Frutos	Fruto
	Gemas	Gema
	Genótipos	Genotipo
	Híbridos	Híbrido
	Laranjeiras	Laranjeira USE Laranja
	Lepidópteros	Lepidoptero
	Maçãs	Maca
	Marcadores moleculares	Marcador molecular
	Micronutrientes	Micronutriente USE Microelemento
	Mudas	Muda
	Pereira (título) e Pereiras (resumo)	Pereira USE Pera
	Pitangas	Pitanga
	Práticas culturais	Pratica cultural
	Produtos químicos	Produto quimico
	Regulador de crescimento (Resumo) e Reguladores de crescimento (texto)	Regulador de crescimento
	Tratamentos	Tratamento

Fonte: Elaborado pela autora

O Quadro 14 mostra situações em que o termo do artigo científico não foi atribuído porque no vocabulário controlado se encontra apenas o termo no singular, como em "*agrotoxico*", "*genotipo*", "*clone*", "*carboidrato*", "*carotenoide*", "*cromossomo*", entre outros.

É possível verificar também que variações ocorreram entre as ocorrências nas estruturas "título", "resumo" e "texto" do artigo. Casos como os do termo "*biofilmes*", apresentado no resumo, e "*biofilme*", no texto; "*pereira*" no título e "*pereiras*" no resumo; "*regulador de crescimento*" no resumo e "*reguladores de crescimento*" no texto. Ou seja, o sistema não é capaz de identificar que, em realidade, essas palavras possuem o mesmo conceito, apesar da diferença morfológica.

Outro fator que identificamos está relacionado ao critério estatístico do SISA quando a frequência de um termo é elevada em apenas uma estrutura do artigo:

QUADRO 15 - Fator de interferência na indexação automática (ocorrência de termos de indexação em apenas uma estrutura do texto)

Fator	**Exemplos**		
	Título	**Resumo**	**Texto**
Frequência de ocorrência dos termos em apenas uma estrutura do documento			Temperatura
			Armazenamento
			Vírus
			Produtividade
		Coco	
			Laranja
			Qualidade
		Mercado	
			Umidade
	Processados	Processados	Processamento
			Adubação
			PH
		Fenologia	
			Clima
		Marmeleiro	Polpa Marmelo

Fonte: Elaborado pela autora

O QUADRO 15 mostra que, em muitos casos, o termo relevante para indexação — "*temperatura*", "*armazenamento*", "*vírus*", entre outros — foi apresentado apenas na estrutura texto do artigo. Em outras situações é apresentado também no título e no resumo, mas com

algumas variações que o impedem de ser computado como conceito equivalente. Nesse contexto, o critério estatístico de ponderação do sistema para definir se um termo será atribuído como termo de indexação já impediu que se considerassem as duas formas como um mesmo conceito.

Uma das questões tratadas na indexação automática é a identificação de termos compostos. No experimento, verificamos a importância que a identificação desses componentes tem para a adequada indexação.

QUADRO 16 - Fator de interferência na indexação automática (dificuldade em atribuir termos compostos)

Fator	**Exemplos**	
	Artigo científico	**Vocabulário controlado**
Dificuldade em atribuir termos compostos	In vitro Cultivo in vitro	Cultura in vitro
	Teste Testes	Teste de vigor
	Características Características fenotípicas	Caracteristicas agronomicas
	Doença	Doença de planta
	Trichogramma	Trichogramma SP
	Propagação	Propagacao vegetativa
	Substrato Substratos	Substrato de cultura
	Ambiente Ambientes	Meio ambiente
	Fisiologia	Fisiologia vegetal
	Resposta	Resposta da planta
	Conservação	Conservacao de alimento USE Preservacao de alimento
	Amadurecimento	Amadurecimento USE Maturacao Maturacao tardia
	Melhoramento	Melhoramento Melhoramento genetico vegetal
	Nutrição	Nutricao Nutricao vegetal
	Análise	Analise foliar
	Praga Planta	Praga de planta
	Distribuição Distribuição espacial	Distribuicao geografica

(continua)

(conclusão)

Fator	Exemplos	
Dificuldade em atribuir termos compostos	Distribuição Distribuição espacial	Distribuicao geografica
	Leprose Laranjeira	Leprose citrica
	Características químicas	Composicao quimica
	Indução Brotação	Brotacao induzida
	Taxa respiratória	Respiracao Taxa

Fonte: Elaborado pela autora

Foi verificado na análise dos artigos que, comumente, os autores utilizam o termo composto inicialmente para delimitar o assunto que será tratado no artigo e, ao longo do artigo, além de anáforas, referem-se ao assunto utilizando apenas a primeira unidade lexical que constitui o termo composto. Por exemplo, usa-se, inicialmente, o termo "*propagação vegetativa*" e, ao longo do texto, "*propagação*", empregando-se um recurso do discurso que permite subentender que se trata de "*propagação vegetativa*". Esse caso ocorreu nos termos "*teste de vigor*", "*doença de planta*", "*substrato de cultura*", "*fisiologia vegetal*", "*resposta da planta*", "*melhoramento genético vegetal*", entre outros.

É possível observar que os termos "*distribuição espacial*", "*características químicas*", indicam, respectivamente, os conceitos "*distribuição geográfica*" e "*composição química*" no contexto do assunto tratado no artigo. Porém, o fato de não constarem no vocabulário controlado permite descartá-los como termos de indexação.

Da mesma forma que os conceitos de "*indução*" e "*brotação*", apresentados no artigo, não são representados pelo termo autorizado "*brotação induzida*", do vocabulário controlado, porque apresentam padrões diferentes na sequência de caracteres.

Verificamos, por outro lado, que o artigo apresenta o termo composto "*taxa respiratória*", mas que o vocabulário controlado não o contempla e permite que seja representado por termos simples como "*taxa*" e "*respiração*".

No QUADRO 17 verificamos as limitações impostas pela indexação automática baseada apenas em análise de padrões linguísticos.

Verificamos a interferência da diferença de padrões expressa no uso e na ausência de hífen, de aspas simples e de parênteses, observados em termos como "*porta-enxertos*", "*jambo-vermelho*", "*maracujazeiro-amarelo*", "*cercas-vivas*", "*mosca-das-frutas*" e "*morte-*

precoce", com hífen; e em termos como "*'ponkan'*" e "*'maçã'*", delimitados por aspas simples; além de termos como "(*myrtaceae*)" delimitado por parênteses.

QUADRO 17 - Fator de interferência na indexação automática (diferença entre as estruturas dos termos de indexação)

Fator	**Exemplos**	
Diferença na apresentação entre os termos do artigo e do vocabulário controlado	**Artigo científico**	**Vocabulário controlado**
	Porta-enxerto[1]	Porta enxerto
	Jambo-vermelho	Jambo
	Quebra de dormência	Quebra da dormencia
	Caractere agronômico Caractere	Caracteristicas agronomicas
	Maracujazeiro-amarelo	Maracujazeiro USE Maracuja
	Cercas-vivas	Cerca viva USE Planta para cerca viva
	'Ponkan'	Ponkan
	Passiflora	Passifloracea
	Banana 'Maçã'	Banana Maca
	Vida de prateleira Vida útil pós-colheita	Vida-de-prateleira
	Mosca-das-frutas	Mosca das frutas
	Morte-precoce	Morte precoce
	K e Ca	Potassio e Calcio
	Goiabeira-serrana	Goiaba serrana
	(Myrtaceae)	Myrtaceae

Fonte: Elaborado pela autora

Além disso, foi constatada a interferência do uso de preposição em termos compostos, como no caso do termo "*quebra de dormência*" no artigo e "*quebra da dormencia*" no vocabulário controlado, em que a diferença se encontra nas preposições "*de*" e "*da*".

Em muitas áreas do conhecimento é comum o uso de símbolos convencionais, siglas e abreviaturas. No experimento, verificamos o uso de símbolos de elementos químicos ao invés do uso da forma por extenso "*cálcio*" e "*potássio*", representados no artigo respectivamente pelos símbolos "Ca" e "K", que não constam no vocabulário controlado.

Outro fator que constatamos foi o problema com a identificação de conceitos implícitos:

QUADRO 18 - Fator de interferência na indexação automática (dificuldade em atribuir conceitos implícitos)

Fator	**Exemplos**	
	Artigo científico	**Vocabulário controlado**
Dificuldade em atribuir conceitos implícitos	...armazenado em câmaras frias...	Refrigeracao
	...reprodução...sistema reprodutivo...polinização	Reproducao vegetal
	...temperatura...ar...umidade...iluminação	Climatologia
	...cultivares...características culturais...cultivo...ciclo de produção	Pratica cultural
	...induzir a brotação...estimular a brotação	Brotacao induzida
	...fertilização....nitrogênio	Fertilizante nitrogenado
	...bactéria...inseto...infecção...plantas	Praga de planta
	...comportamento de novas cultivares...	Comportamento de variedade
	...características químicas...características físicas	Propriedade fisico-quimica
	...enxerto...produção de mudas ...enraizamento de estacas...estaquia de ramos...	Propagacao vegetativa
	...danos mecânicos...dano físico...dano externo	Dano mecanico
	...pós-colheita...embalagem	Conservacao de alimento USE Preservacao de alimento
	...lepidópteros minadores...	Lagarta minadora
	...efeito dos resíduos...efeito tóxico...efeito desses agrotóxicos	Efeito residual
	...crescimento vegetativo...desenvolvimento vegetativo	Propagacao vegetativa
	...variabilidade genética...variabilidade intra-específica	Variacao genetica
	...teores foliares...avaliação nutricional	Analise foliar

Fonte: Elaborado pela autora

A flexibilidade da linguagem oferece recursos linguísticos que permitem descrever um assunto de diversas formas. Nesse sentido, alguns conceitos podem estar implícitos nos textos e sua identificação se torna difícil no processo de indexação automática. No QUADRO 18

verificamos que os trechos do artigo indicam um determinado termo de indexação representado no vocabulário controlado, que, entretanto, não é atribuído, porque nem mesmo é mencionado no artigo.

Uma das questões que surge quando se discute a indexação automática é a garantia do aspecto semântico, uma vez que caracteriza a qualidade atribuída à indexação. Dessa forma, identificamos alguns termos atribuídos pelo SISA que não correspondem ao significado que o termo representa no vocabulário controlado.

QUADRO 19 - Fator de interferência na indexação automática (diferença semântica nos termos de indexação)

Fator	**Exemplos**		
	Artigo científico	**Vocabulário controlado**	**Termo atribuído por SISA**
Diferença semântica nos termos de indexação	Pelo (preposição)	Pelo (pelagem)	Pelo
	Pato Branco (município)	Pato (ave)	Pato
	Mato Grosso do Sul (Estado)	Mato (vegetação)	Mato
	Sr. Pinto (variedade de laranjeira)	Pinto (ave)	Pinto
	Pêra (variedade de laranjeira)	Pera (fruto da pereira)	Pera
	Capão Bonito (município)	Bonito (espécie de peixe)	Bonito
	Rio Grande do Norte (Estado)	Rio (canal fluvial)	Rio
	Primor Amoreira (variedade de mangueira)	Amoreira USE Amora	Amora
	Rosa (variedade de mangueira)	Rosa (flor da roseira)	Rosa

Fonte: Elaborado pela autora

O QUADRO 19 mostra os termos de indexação atribuídos pelo SISA ao considerar palavras isoladas que em realidade, são unidades lexicais que constituem um termo composto presente no artigo. No entanto, o termo composto apresenta significado totalmente diferente do representado no termo do vocabulário controlado. É possível constatar, por exemplo, que a palavra que constitui o nome de um município como "*Capão Bonito*" é identificada no vocabulário, mas possui outro conceito, visto que o vocabulário entende "*Bonito*" como uma espécie de peixe. No entanto, o sistema automático não é capaz de distinguir o contexto semântico de cada termo e os atribui como se fossem conceitos equivalentes.

Um dos fatores que tem influência sobre os índices de consistência na indexação é a quantidade de termos atribuídos. Constatamos que o SISA proporciona um número elevado de termos de indexação em relação à indexação manual. Nesse sentido, um dos motivos que explica o elevado número de termos é a atribuição dos termos gerais e também dos específicos.

QUADRO 20 - Fator de interferência na indexação automática (atribuição automática de termo geral e de termo específico)

Fator	**Exemplos**	
	Termo geral	**Termo específico**
Atribuição automática de termo geral e de termo específico	Clima	Clima temperado
	Universidade	Universidade Federal
	Eugenia	Eugenia Involucrata
	Areia	Areia fina
	Óleo	Óleo mineral
	Distúrbio	Distúrbio fisiológico
	Adubação	Adubação verde
	Casca	Casca de arroz
	Iluminação	Iluminação artificial
	Umidade	Umidade relativa
	Planta	Planta hospedeira
	Mercado	Mercado atacadista
	Laranja	Laranja pêra
	Deficiência	Deficiência hídrica
	Profundidade	Profundidade de semeadura
	Ácido	Ácido indolbútirico
	Latossolo	Latossolo amarelo
	Ácido	Ácido bórico
	Fisiologia	Fisiologia vegetal
	Microscopia	Microscopia eletrônica

Fonte: Elaborado pela autora

Atribuir o termo geral é interessante para conferir à indexação maior exaustividade na busca de informação, mas também interfere na precisão. Em análise comparativa com a indexação manual, foi possível notar que a especificidade norteia a política de indexação da BINAGRI. O SISA, por outro lado, indexa o termo geral sem considerar que o artigo trata em realidade apenas do termo específico constituído por composição.

Além desse fator, o que tem gerado uma grande quantidade de termos propostos pelo SISA é a indexação de termos relacionados à metodologia da pesquisa.

QUADRO 21 - Fator de interferência na indexação automática (atribuição de termos relacionados à metodologia da pesquisa)

Fator	**Exemplos**
Atribuição de termos relacionados à metodologia da pesquisa	**Termos**
	Trabalho
	Método estatístico
	Análise
	Método
	Pesquisa
	Areia fina
	Papel
	Estação experimental
	Escola
	Agar
	Tecnologia
	Mandioca
	Laboratório
	Vinagre

Fonte: Elaborado pela autora

Entre os termos analisados, verificamos que a palavra "*trabalho*" é atribuída a muitos artigos científicos porque se apresenta em frases como "*(...) este trabalho tem o objetivo de(...)*", assim como os termos "*pesquisa*"*, "análise*" e "*método*", geralmente presentes no resumo e na introdução do texto, contemplando o critério de ponderação para atribuição pelo SISA.

Materiais como "*areia fina*", "*papel*", "*Agar"* e "*vinagre*", aplicados nos experimentos na área agrícola, foram atribuídos por condição idêntica à circunstância mencionada acima.

Termos como "*estação experimental*", "*escola*", "*tecnologia*", "*mandioca*" e "*laboratório*" foram identificados em nomes de instituições em que o experimento da pesquisa foi realizado. Em casos como a atribuição dos termos "*tecnologia*" e "*mandioca*", fica evidente também a diferença do significado expresso pelo conceito no artigo e no vocabulário controlado.

Ademais, verificamos que, nos artigos científicos, os autores se referem tanto ao fruto, por exemplo, "*maçã*", como também ao vegetal que produz a fruta, "*macieira*". Nesse caso, o vocabulário controlado contempla a equivalência *Macieira USE Maca*. Em alguns casos, porém, verificamos não haver essa equivalência, o que comprometeu a atribuição por indexação automática.

QUADRO 22 - Fator de interferência na indexação automática (relação de equivalência omitida)

Fator	**Exemplos**	
Relação de equivalência omitida	**Artigo científico**	**Vocabulário controlado**
	Lichieira	Lichia
	Aceroleira	Acerola
	Pequizeiro	Pequi
	Pereira	Pera

Fonte: Elaborado pela autora

Cabe ressaltar que o estabelecimento de equivalência do vocabulário controlado é um recurso que permite que muitos termos de indexação sejam atribuídos e sua ausência nos casos citados ocasionou inconsistência na indexação.

Em síntese, os fatores intervenientes nos índices de consistência na indexação estão relacionados às diferenças linguísticas de âmbito morfológico, sintático e semântico entre os termos do vocabulário controlado e os termos dos artigos científicos.

Foram expostos exemplos claros que mostram diferenças entre termos no singular e no plural, constatando que, na maioria das vezes, os autores de artigos científicos se valem do plural para tratar do assunto do artigo, ao passo que o vocabulário controlado, em geral, apresenta o termo no singular.

Existem diversas formas de expor um assunto e, nos artigos científicos, em alguns casos, os termos de indexação são tratados com mais frequência na estrutura texto, principalmente na introdução, em que é esclarecida a proposta do estudo em questão. Por isso, alguns termos podem não ser atribuídos pelo sistema, se forem considerados seus critérios de ponderação dos termos em estruturas específicas do artigo, apesar do sistema possuir critérios de ponderação quando um termo tem frequência elevada apenas no texto[32].

[32] Se o termo candidato a descritor aparece no título, resumo e texto, apresenta-se ao indexador para sua possível incorporação como termo de indexação. Se um termo candidato a descritor aparece no texto dez vezes ou mais,

Um dos fatores intensamente verificado foi o problema em atribuir termos compostos. Constatamos que a sua omissão oferece prejuízos na indexação, na medida em que representam conceitos específicos que não podem simplesmente ser representados por seu fragmento. O fragmento de um termo composto representa um conceito diferente e descontextualiza a indexação de um documento. É necessário observar que não é possível identificar automaticamente os termos compostos por simples critérios de frequência, visto que os autores se valem da primeira unidade lexical do termo composto para, ao longo do artigo científico, referirem-se ao termo composto.

Sendo assim, verificamos que o SISA tem-se apoiado na comparação de padrões de sequência de caracteres, o que acarreta problemas na indexação. Todos os fatores examinados, de certa forma, estão relacionados às consequências que esse critério de análise automática ocasiona. Essa circunstância pode ser visualizada na interferência sobre a atribuição de termos de indexação causada pelo uso de símbolos, parênteses, apóstrofos, aspas e pelas diferenças de preposições entre os termos do artigo e do vocabulário controlado.

Verificamos também essa circunstância quando o sistema encontra dificuldade em atribuir conceitos implícitos nos artigos e quando atribui um termo de indexação que possui conceito distinto do representado por um termo apresentado no vocabulário controlado. Distinguir os aspectos semânticos é uma característica fundamental para garantir a apropriada recuperação da informação.

6.2 Índices de exaustividade e precisão na recuperação da informação

Apresentamos na TAB. 2 os índices médios de exaustividade e precisão obtidos nas buscas realizadas nas bases de dados BDSISA e BDBINAGRI e, em seguida, cada um dos fatores intervenientes na recuperação dos artigos científicos.

---

além de aparecer em oito parágrafos diferentes ou mais, e não está incluído em nenhum dos termos propostos, apresenta-se como termo candidato a termo de indexação.

TABELA 2 - Índice médio de exaustividade e precisão na recuperação da informação nas bases de dados

	BDSISA		BDBINAGRI	
	Exaustividade (%)	**Precisão (%)**	**Exaustividade (%)**	**Precisão (%)**
Soma total	2651	2850	3368	3672
Índice médio	2651/50= 53,02%	2850/50= 57%	3368/50= 67,36%	3672/50= 73,44%

Fonte: Elaborada pela autora

Das cinquenta buscas de pesquisa realizadas, calcularam-se os índices médios de exaustividade e de precisão na recuperação da informação em ambas as bases de dados. Na base de dados BDSISA obteve-se o índice médio de 53,02% de exaustividade, enquanto na BDBINAGRI obteve-se 67,36%. Com relação à precisão na recuperação da informação, o índice da base de dados BDSISA também foi inferior ao da base de dados BDBINAGRI, com 57% contra 73,44%.

6.2.1 Fatores intervenientes nos índices de exaustividade e de precisão na recuperação da informação

De modo geral, o fato do sistema não ter atribuído diversos termos de indexação foi decisivo no desempenho da recuperação da informação, no sentido de que, por não constar como ponto de acesso ao assunto, o artigo não foi recuperado. Nesse sentido, os quadros a seguir mostram os fatores e os exemplos verificados, indicando os artigos que não foram recuperados:

QUADRO 23 - Fator de interferência na recuperação da informação (flexão de número nos termos de indexação)

Fatores	**INDEXAÇÃO**		**RECUPERAÇÃO DA INFORMAÇÃO**		**Comentário**
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Termos no singular e no plural	Bananas Bananeiras	Banana Bananeira USE Banana	Adubação E Banana	Exaustividade	O artigo 99 não foi recuperado porque o termo “Banana” não foi atribuído na indexação
	Pitangas	Pitanga	Armazenamento E Pitanga	Exaustividade e precisão	O artigo 12 não foi recuperado porque o termo “Pitanga” não foi atribuído na indexação

(continua)

(conclusão)

	INDEXAÇÃO		RECUPERAÇÃO DA INFORMAÇÃO		Comentário
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Termos no singular e no plural	Sementes	Semente	Germinação E Semente	Exaustividade e precisão	Os artigos 7 e 62 não foram recuperados porque o termo "Semente" não foi atribuído na indexação
			Semente E Atemóia	Exaustividade e precisão	O artigo 28 não foi recuperado porque o termo "Semente" não foi atribuído na indexação
	Macieiras	Macieira USE Maca	Quebra da dormência E Maçã	Exaustividade e precisão	Artigo 78 não foi recuperado porque o termo "Maçã" não foi atribuído na indexação
	Agrotóxicos	Agrotoxico	Agrotóxico E Fruta cítrica	Exaustividade e precisão	Os artigos 46 e 86 não foram recuperados porque o termo "Agrotóxico" não foi atribuído na indexação
	Biofilmes (resumo) Biofilme (texto)	Biofilme	Biofilme E Pós-colheita E Manga	Exaustividade e precisão	O artigo 82 não foi recuperado porque o termo "Biofilme" não foi atribuído na indexação
	Maçãs (título, resumo) Maçã (texto) Macieiras	Maca Macieira USE Maca	Maçã	Exaustividade	O artigo 13 não foi recuperado porque o termo "Maçã" não foi atribuído na indexação
	Gemas	Gema	Gema E Pêra	Exaustividade e precisão	O artigo 1 não foi recuperado porque o termo "Gema" não foi atribuído na indexação
	Substratos	Substrato	Substrato E Enraizamento	Exaustividade e precisão	Os artigos 77, 95 e 97 não foram recuperados porque o termo "Substrato" não foi atribuído na indexação

Fonte: Elaborado pela autora

QUADRO 24 - Fator de interferência na recuperação da informação (ocorrência de termos de indexação em apenas uma estrutura do texto)

Fatores	**INDEXAÇÃO**		**RECUPERAÇÃO DA INFORMAÇÃO**		**Comentário**
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Frequência de ocorrência dos termos em apenas uma estrutura do documento	Pós-colheita (texto)	Pos-colheita	Armazenamento E Pós-colheita	Exaustividade	O artigo 42 não foi recuperado porque o termo "Pós-colheita" não foi atribuído na indexação
			Pós-colheita	Exaustividade	O artigo 42 não foi recuperado porque o termo "Pós-colheita" não foi atribuído na indexação

QUADRO 25- Fator de interferência na recuperação da informação (dificuldade em atribuir termos compostos)

Fatores	INDEXAÇÃO		RECUPERAÇÃO DA INFORMAÇÃO		Comentário
	Artigo Científico	Vocabulário Controlado	Estratégia de Busca	Interferência Exaustividade/Precisão	
Dificuldade em atribuir termos compostos	Análise	Analise Analise Foliar	Análise foliar E Banana	Exaustividade	O artigo 25 não foi recuperado porque o termo "Análise foliar" não foi atribuído na indexação
	Conservação	Conservacao de alimento USE Preservacao de alimento	Preservação de alimento E Pós-colheita	Exaustividade e precisão	Os artigos 12, 14, 41, 43 e 82 não foram recuperados porque o termo "Preservação de alimento" não foi atribuído na indexação
	Colheita	Pos-colheita	Preservação de alimento E Pós-colheita	Exaustividade e precisão	O artigo 42 não foi recuperado porque o termo "Pós-colheita" não foi atribuído na indexação
	Propagação	Propagacao vegetativa	Propagação vegetativa E Maracujá	Exaustividade	O artigo 59 não foi recuperado porque o termo "Propagação vegetativa" não foi atribuído na indexação
	Teste Testes	Teste de vigor	Teste de vigor E Semente E Mangaba	Exaustividade e precisão	O artigo 8 não foi recuperado porque o termo "Teste de vigor" não foi atribuído na indexação
	Variedade	Variedade resistente	Variedade resistente E Banana	Exaustividade e precisão	O artigo 34 não foi recuperado porque o termo "Variedade resistente" não foi atribuído na indexação
			Variedade resistente E Uva	Exaustividade e precisão	Artigo 56 não foi recuperado porque o termo "Variedade resistente" não foi atribuído na indexação
	Propagação	Propagacao vegetativa	Propagação vegetativa	Exaustividade	Os artigos 59, 60 e 95 não foram recuperados porque o termo "Propagação vegetativa" não foi atribuído na indexação
	Conservação	Preservacao de alimento Conservaçca de alimento USE Preservaçcao de alimento	Preservação de alimento	Exaustividade e precisão	Os artigos 12, 14, 41 e 43 não foram recuperados porque o termo "Preservação de alimento" não foi atribuído na indexação

(continua)

(conclusão)

	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
	Doença	Doenca de planta	Doença de planta	Exaustividade e precisão	Os artigos 32 e 35 não foram recuperados porque o termo "Doença de planta" não foi atribuído na indexação
	Análise	Analise foliar	Análise foliar	Exaustividade e precisão	O artigo 25 não foi recuperado porque o termo "Análise foliar" não foi atribuído na indexação
	Melhoramento	Melhoramento Melhoramento genetico vegetal	Melhoramento genético vegetal E Uva	Exaustividade e precisão	O artigo 24 não foi recuperado porque o termo "Melhoramento genético vegetal" não foi atribuído na indexação

Fonte: Elaborado pela autora

QUADRO 26 - Fator de interferência na recuperação da informação (diferença entre estruturas dos termos de indexação)

Fatores	**INDEXAÇÃO**		**RECUPERAÇÃO DA INFORMAÇÃO**		**Comentário**
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Diferença na apresentação entre os termos do artigo e do vocabulário controlado	Porta-enxerto	Porta enxerto	Porta enxerto E Pêssego	Exaustividade e precisão	Os artigos 51 e 64 não foram recuperados porque o termo "Porta enxerto" não foi atribuído na indexação
	Jambo-vermelho	Jambo	Semente E Jambo	Exaustividade e precisão	O artigo 27 não foi recuperado porque o termo "Jambo" não foi atribuído na indexação
	Quebra de dormência	Quebra da dormencia	Quebra da dormência E Maçã	Exaustividade e precisão	Os artigos 30 e 78 não foram recuperados porque o termo "Quebra da dormência" não foi atribuído na indexação

Fonte: Elaborada pela autora

QUADRO 27 - Fator de interferência na recuperação da informação (dificuldade em atribuir conceitos implícitos)

Fatores	INDEXAÇÃO		RECUPERAÇÃO DA INFORMAÇÃO		Comentário
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Dificuldade em atribuir conceitos implícitos	Produção	Producao Produtividade	Poda E Produtividade	Exaustividade e precisão	O artigo 73 não foi recuperado porque o termo "Produtividade" não foi atribuído na indexação
	Manejo integrado de pragas...infestação pelo pulgão....	Controle biologico	Controle biológico E Fruta cítrica	Exaustividade e precisão	O artigo 44 não foi recuperado porque o termo "Controle biológico" não foi atribuído na indexação
	Produção Produtividade	Producao Produtividade	Produtividade E Banana	Exaustividade e precisão	O artigo 23 não foi recuperado porque o termo "Produtividade" não foi atribuído na indexação
	...enxerto...produção de mudas ...enraizamento de estacas...estaquia de ramos...	Propagacao vegetativa	Propagação vegetativa	Exaustividade	Os artigos 77 e 97 não foram recuperados porque o termo "Propagação vegetativa" não foi atribuído na indexação

Fonte: Elaborado pela autora

QUADRO 28 - Fator de interferência na recuperação da informação (artigos irrelevantes recuperados)

Fatores	INDEXAÇÃO		RECUPERAÇÃO DA INFORMAÇÃO		Comentário
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Recuperação de artigos científicos além do estabelecido como relevante para a busca	---	--	Maturação E Pós-colheita	Precisão	Os artigos 13, 14, 41 e 69 foram recuperados. Na maioria dos casos verificamos que o termo "maturação" ou o termo "pós-colheita" tratavam-se de um assunto secundário
	---	---	Micropropagação	Precisão	Os artigos 60, 76 e 95 foram recuperados
	---	---	Armazenamento E Pós-colheita	Precisão	Os artigos 7, 14, 41 e 69 foram recuperados
	---	---	Ácido giberélico E Floração	Precisão	O artigo 73 foi recuperado

Fonte: Elaborado pela autora

Em menor número, ocorreram casos em que se recuperaram artigos que não constavam como relevantes para determinada busca. Na maioria dos casos, os termos utilizados na estratégia de busca foram atribuídos na indexação automática, mas a análise do artigo científico revelou que se tratavam de assuntos secundários.

QUADRO 29 - Fator de interferência na indexação automática (relação de equivalência omitida no vocabulário controlado)

Fatores	**INDEXAÇÃO**		**RECUPERAÇÃO DA INFORMAÇÃO**		**Comentário**
	Artigo Científico	**Vocabulário Controlado**	**Estratégia de Busca**	**Interferência Exaustividade/Precisão**	
Relação de equivalência omitida	Lichieira	Lichia	Frutificação E Lichia	Exaustividade e precisão	Os artigos 4 e 5 não foram recuperados porque o termo “Lichia” não foi atribuído na indexação
	Pereira	Pera	Gema E Pêra	Exaustividade e precisão	O artigo 9 não foi recuperado porque o termo “Pêra” não foi atribuído na indexação
	Aceroleira	Acerola	Estaca E Acerola	Exaustividade e precisão	O artigo 19 não foi recuperado porque o termo “Acerola” não foi atribuído na indexação

Fonte: Elaborado pela autora

Os quadros expõem como os fatores intervenientes na indexação afetaram a recuperação da informação. Na coluna “indexação”, os exemplos indicam que não houve coincidência entre o termo do artigo científico e o do vocabulário controlado, o que justifica o fato de não ter sido atribuído como termo de indexação. O exemplo da indexação é relacionado à estratégia de busca, que não foi contemplada, haja vista a interferência nas características de exaustividade ou nas de precisão ou em ambas durante a recuperação da informação.

Os fatores “diferenças semântica nos termos de indexação”, “atribuição automática de termo geral e de termo específico” e “atribuição de termos relacionados à metodologia da pesquisa” não foram identificados na análise da recuperação da informação, mas podem interferir e ocasionar problemas relacionados à precisão na busca de informação.

O fator “diferenças semântica nos termos de indexação” pode acarretar a recuperação de documentos totalmente irrelevantes para a busca solicitada pelo usuário, ao atribuir um

termo de indexação que possui significado diferente do determinado pelo vocabulário controlado.

O fator “atribuição automática de termo geral e de termo específico” pode oferecer uma elevada exaustividade na busca, mas os resultados da busca exigem do usuário uma análise minuciosa e dispêndio de tempo.

O fator “atribuição de termos relacionados à metodologia da pesquisa” também pode acarretar a falta de precisão nas buscas, ao oferecer como resultado de busca documentos que não tratam especificamente sobre determinado assunto, mas apenas aplicam um material, tipo de análise para desenvolver a pesquisa ou, como verificamos, indicam uma parte do nome da instituição em que a pesquisa foi desenvolvida.

Nesse contexto, os exemplos apresentados nos quadros indicam como a omissão de termos de indexação relevantes pode desvirtuar a qualidade da indexação e a recuperação da informação. Assim como tanto a atribuição em excesso como a atribuição de assuntos secundários conduzem à exaustividade e até mesmo a ruídos na recuperação da informação.

Portanto, a partir da análise da atuação do sistema SISA é possível verificar que os critérios de análise automática do sistema com relação à análise linguística não têm sido suficientes para aplicar o vocabulário controlado com confiabilidade. O sistema não possui o tratamento morfossintático para tratar das situações de diferenças de flexão de número e gênero, uso de símbolos, preposições, etc. nas estruturas textuais do documento durante a indexação automática, o que tem acarretado algumas limitações no processo de indexação.

Além disso, o vocabulário controlado aplicado no sistema foi elaborado para aplicação em indexação realizada por humanos e, portanto, considera que seu uso seja realizado por processos de interpretação dos conceitos. Sua aplicação na indexação automática enfrenta problemas na transposição dos termos identificados no texto para o vocabulário controlado. Nesse processo, o vocabulário controlado acaba condicionando os resultados de indexação, na medida em que são considerados termos de indexação apenas os identificados no vocabulário controlado. Os dados do experimento revelam esse aspecto ao omitir termos de indexação por não constarem da forma apresentada no vocabulário controlado, assim como nos casos de conceitos implícitos nos textos.

Constatamos que a aplicação de vocabulário controlado na indexação automática deve considerar aspectos particulares apresentados pelos critérios empregados para atribuir os termos de indexação. Tais aspectos se referem à elaboração de uma rede de conceitos por meio da qual o sistema atue para considerar: conceitos implícitos; precisão conceitual em

palavras polissêmicas; e identificação de termo geral e específico, para atender a decisão de uma política de indexação e, enfim, para permitir uma abordagem contextual na indexação automática.

É importante ressaltar que o experimento foi realizado com uma amostra de 100 artigos científicos, uma parcela pequena comparada à quantidade de artigos de bases de dados em geral. Entretanto, a partir desse contexto constataram-se algumas circunstâncias que permitem levantar questões discutidas na literatura.

Portanto, apresentamos no próximo capítulo a análise e a discussão dos dados qualitativos e quantitativos obtidos nesta pesquisa e em pesquisa anterior (NARUKAWA, GIL LEIVA, FUJITA, 2009), para que a fundamentação teórica contribua com a análise das situações identificadas nos experimentos e nos permita levantar alguns aspectos a serem considerados na atuação de vocabulários controlados em indexação automática.

7 IMPLICAÇÕES SOBRE O USO DOS VOCABULÁRIOS CONTROLADOS NO PROCESSO DE INDEXAÇÃO AUTOMÁTICA

Este capítulo apresenta a discussão dos resultados da pesquisa a partir das constatações dos experimentos em que aplicamos os vocabulários controlados DeCS e ThesAgro no sistema de indexação automática SISA. Resgatamos os dados obtidos e analisamos as implicações de cada fator interveniente na indexação sob a perspectiva das discussões apontadas no referencial teórico para, finalmente, apresentar alguns aspectos que merecem maior atenção na adaptação do vocabulário controlado em indexação automática.

A pesquisa de Narukawa, Gil Leiva e Fujita (2009) desenvolveu-se a partir da aplicação do vocabulário controlado DeCS no sistema SISA com 100 artigos científicos da área de odontologia. Ao analisar os dados, verificamos que o vocabulário empregado no processo de indexação automática tem elevada interferência sobre os resultados do processo, justamente por condicionar a atribuição de termos de indexação.

O experimento realizado a partir do uso do vocabulário ThesAgro no sistema SISA foi aplicado com 100 artigos científicos da área agrícola e seus resultados confirmam os fatores que interferiram no processo de indexação com o DeCS. Identificamos os mesmos fatores, com exceção do fator “Relação de equivalência omitida”.

A análise dos dados foi realizada por meio do exame de cada fator interveniente na consistência da indexação e que interferiu também na recuperação da informação nas bases de dados em ambas as pesquisas. Buscamos analisar cada fator para justificar os índices obtidos e interpretá-los sob a perspectiva das alternativas apresentadas no referencial teórico.

7.1 Análise das avaliações intrínseca e extrínseca

Na TAB. 3 são sistematizados os índices de consistência na indexação, o de exaustividade e o de precisão na recuperação da informação de ambas as pesquisas.

TABELA 3 - Índice médio de consistência na indexação, de exaustividade e de precisão na recuperação da informação

	Aplicação do ThesAgro		Aplicação do Decs	
	SISA	**BINAGRI**	**SISA**	**BIREME**
Consistência na indexação	19,30%		23,25%	
Exaustividade na recuperação da informação	53,02%	67,36%	35, 72%	77, 04%
Precisão na recuperação da informação	57%	73,44%	40, 92%	78, 69%

Fonte: Elaborado pela autora

Apesar da sutil diferença entre os índices de consistência nas pesquisas, os fatores identificados na aplicação do ThesAgro apenas ratificaram os identificados em pesquisa anterior, confirmando alguns problemas enfrentados na indexação automática. Os índices de recuperação da informação na base de dados BDSISA com uso do ThesAgro apresentaram índice mais alto em relação à pesquisa na base de dados BDSISA com uso do DeCS, ainda que a aplicação do DeCS tenha apresentado melhor índice de consistência na indexação.

7.1.1 Análise dos fatores intervenientes na consistência da indexação

A partir da análise da consistência entre a aplicação do ThesAgro e a aplicação do DeCS na indexação automática, identificamos que o principal motivo que impediu o sistema SISA de atribuir muitos termos de indexação relevantes deveu-se à diferença entre o termo utilizado no artigo e o termo do vocabulário controlado, problematizada pelos seguintes fatores:

- ✓ Termos no singular e no plural;
- ✓ Frequência de ocorrência dos termos em apenas uma estrutura do documento;
- ✓ Dificuldade em atribuir termos compostos;
- ✓ Diferença na apresentação dos termos do artigo e do vocabulário controlado;
- ✓ Dificuldade em atribuir conceitos implícitos;
- ✓ Diferença semântica nos termos de indexação;
- ✓ Atribuição automática de termo geral e de termo específico;

✓ Atribuição de termos relacionados à metodologia da pesquisa.

Além desses, há o fator "relação de equivalência omitida", identificado apenas no experimento com aplicação do ThesAgro.

a) Termos no singular e no plural

Ocorreram diferenças entre os termos do artigo científico e os do vocabulário controlado no que se refere à flexão de número.

QUADRO 30 - Fator interveniente na aplicação do ThesAgro e do DeCS (termos no singular e no plural)

Termos no singular e no plural			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
Agrotóxicos	Agrotoxico	Ameloblastomas (título e resumo) Ameloblastoma (texto)	Ameloblastoma
Biofilmes (resumo) e Biofilme (texto)	Biofilme	Chupeta	Chupetas
Carboidratos	Carboidrato	Cimento de ionômero de vidro	Cimentos de ionômeros de vidro
Carotenóides	Carotenoide	Dentadura	Dentaduras
Clones	Clone	Doença periodontal	Doenças periodontais
Compostos fenólicos	Composto fenolico	Laser	Lasers
Cromossomos	Cromossomo	Neurilemomas	Neurilemoma
Custos de Produção	Custo de Producao	Osteomas	Osteoma
Fitorreguladores	Fitorregulador USE Regulador de crescimento	Questionário	Questionarios
Frutos	Fruto	Refrigerante	Refrigerantes
Gemas	Gema	Resina Composta	Resinas Compostas
Genótipos	Genotipo	Resistências à tração	Resistência a tracao
Híbridos	Hibrido	Tumor odontogênico	Tumores odontogenicos
Laranjeiras	Laranjeira USE Laranja	Vasoconstritor (título e resumo) Vasoconstritores (texto)	Vasoconstritores
Lepidópteros	Lepidoptero		
Maçãs	Maca		
Marcadores moleculares	Marcador molecular		
Micronutrientes	Micronutriente USE Microelemento		
Mudas	Muda		
Pereira (título) e Pereiras (resumo)	Pereira USE Pera		

(continua)

(conclusão)

Termos no singular e no plural			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
Pitangas	Pitanga		
Práticas culturais	Pratica cultural		
Produtos químicos	Produto quimico		
Regulador de crescimento (resumo) e Reguladores de crescimento (texto)	Regulador de crescimento		
Tratamentos	Tratamento		

Fonte: Elaborado pela autora

No QUADRO 30 identificamos que as flexões de número interferiram no processo de atribuição de termos de indexação. No entanto, é possível verificar que, no vocabulário controlado ThesAgro, a maioria dos descritores é apresentada no singular, ao contrário do DeCS, que apresenta descritores no plural. Segundo a norma para construção de vocabulários controlados ANSI/NISO Z39.19-2005, os conceitos contáveis devem ser representados no plural, enquanto os não contáveis e os conceitos abstratos são apresentados no singular. Quando um mesmo termo designa uma operação e o produto da mesma, a operação é representada no singular e o seu produto no plural, qualificando o processo com a expressão entre parênteses.

Para o processo de indexação automática, a interferência se concretiza na diferença de padrões entre os termos do artigo e os do vocabulário controlado. Dessa forma, dentre os métodos de indexação automática que permitem tratar essas diferenças verificamos o processo de lematização. A lematização consiste em um processo de redução das palavras, ou de conjunto de palavras, à sua raiz. Assim, o sistema automaticamente identifica que, mesmo possuindo flexão de número, essas palavras possuem o mesmo conceito.

Contudo, existem controvérsias sobre o uso do processo de lematização. Como verificamos na análise dos sistemas de indexação automática, muitos aplicam a lematização na etapa inicial em que o analisador morfológico atua na normalização linguística do texto.

Autores como Câmara Júnior (2007) e Anderson & Pérez-Carballo (2001) ressaltam as complicações que ocorrem nesse processo. É necessário lembrar que a eliminação de “s” não é uma regra amplamente aplicável para distinguir a flexão de número. Além disso, estabelecer

a relação entre a forma da palavra e o significado pode conduzir a erros, pois existem os casos de polissemia, assim como existe um conjunto de palavras que são derivadas da mesma raiz, mas aplicadas em um contexto particular que distingue o seu significado.

b) Frequência de ocorrência dos termos em apenas uma estrutura do documento

Termos de indexação relevantes deixaram de ser atribuídos porque se apresentam em somente uma estrutura do artigo científico.

QUADRO 31 - Fator interveniente na aplicação do ThesAgro e DeCS (frequência de ocorrência dos termos em apenas uma estrutura do documento)

Frequência de ocorrência dos termos em apenas uma estrutura do documento					
Aplicação do ThesAgro no SISA			**Aplicação do DeCS no SISA**		
Título	**Resumo**	**Texto**	**Título**	**Resumo**	**Texto**
		Temperatura			Cisto Radicular
		Armazenamento			Radiografia panorâmica
		Vírus			Sistema Estomatognático
		Produtividade			Cisto dentígero
	Coco				Oncogenes
		Laranja			Genes Supressores de Tumor
		Qualidade			Doenças Periodontais
	Mercado		Saúde	Saúde	Saúde Bucal
		Umidade			Calor
Processados	Processados	Processamento			Mucosa bucal
		Adubação		Cálculos nas glândulas salivares	Cálculos salivares
		PH			Osteoblastoma
	Fenologia				Hipertensão
		Clima			Desinfecção
		Polpa			Cisto radicular
	Marmeleiro	Marmelo			Fatores de risco
					Palato
					Fluoretos
					Fluorose dentária
					Comportamento
					Microscopia eletrônica de varredura

Fonte: Elaborado pela autora

No QUADRO 31 verificamos que as estruturas “título” e “resumo” não apresentam explicitamente o termo de indexação. O assunto é tratado com profundidade na estrutura “texto” do artigo, principalmente na introdução, parte em que o autor expõe o foco da pesquisa indicando a proposta e os objetivos da pesquisa.

Investigações voltadas à exploração das estruturas textuais de artigos científicos podem contribuir para identificar com precisão os conceitos a serem utilizados para a representação da informação. A identificação de conceitos a partir da exploração da estrutura textual de artigos científicos durante a leitura na indexação manual foi investigada por Fujita e Rubi (2006). Para a indexação automática também se torna importante, por evidenciar a parte dos documentos em que se concentram os principais conceitos e onde o sistema poderá atuar na análise automática.

Entre pesquisas que apontam a análise das estruturas textuais “título” e “resumo” como fontes de informação suficientes ou insuficientes para identificar os conceitos para indexação, Gil Leiva e Rodríguez Muñoz (1997) constataram que, na indexação automática, a estrutura “texto” também merece atenção especial por apresentar conceitos relevantes para indexação que não são expressos no título e no resumo. A pesquisa de Gil Leiva e Rodríguez Muñoz (1997) aponta que os títulos e resumos apresentaram, em artigos da área de Biblioteconomia e Documentação, 47,2% dos termos de indexação e, apenas no texto, foi identificado 24,7% dos termos de indexação, correspondendo a uma margem considerável e que deve ser tratada como uma fonte importante para análise automática.

O SISA analisa também os termos de indexação que ocorrem apenas na estrutura “texto”, mas a atribuição é realizada se o termo aparece no texto dez vezes ou mais, além de aparecer em oito parágrafos diferentes ou mais, e se não está incluído em nenhum dos termos propostos. Nos exemplos apresentados no QUADRO 31, a ocorrência dos termos não alcançou esse patamar apesar de serem termos relevantes para indexação. Essa situação merece uma reflexão sobre a possibilidade de inclusão de outros critérios de ponderação de termos de indexação.

Associadas ao fator “termo no singular e no plural”, as diferenças entre termos no plural e no singular ocorreram entre as estruturas textuais “título”, “resumo” e “texto”. Ou seja, o sistema não atribui determinado conceito recorrente em várias estruturas textuais por considerá-los conceitos distintos por causa das flexões de número.

c) Dificuldade em atribuir termos compostos

Termos constituídos por mais de uma unidade lexical não foram atribuídos porque apresentam diferenças entre os termos do artigo científico e do vocabulário controlado.

QUADRO 32 - Fator interveniente na aplicação do ThesAgro e DeCS (dificuldade em atribuir termos compostos)

Dificuldade em atribuir termos compostos			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
In vitro Cultivo in vitro	Cultura in vitro	Adenoma pleomórfico	Adenoma pleomorfo
Teste Testes	Teste de vigor	Diagnóstico	Diagnostico por imagem
Características Características fenotípicas	Caracteristicas agronomicas	Tomografia computadorizada	Tomografia Computadorizada por raio X Tomografia computadorizada de emissao
Doença	Doenca de planta	Ressonância Magnética	Espectroscopia de ressonancia magnetica
Trichogramma	Trichogramma SP	Crescimento gengival	Crescimento excessivo da gengiva
Propagação	Propagacao vegetativa	Interpretação imaginológica	Interpretacao de imagem assistida por computador
Substrato Substratos	Substrato de cultura	Glândulas salivares	Glandulas salivares menores
Ambiente Ambientes	Meio ambiente	Diagnóstico	Diagnostico diferencial
Fisiologia	Fisiologia vegetal	Resinas	Resinas compostas
Resposta	Resposta da planta	Restaurações	Restauracoes intracoronarias
Conservação	Conservacao de alimento USE Preservacao de alimento	Imunoglobulina A	Imunoglobulina A secretora
Amadurecimento	Amadurecimento USE Maturacao Maturacao tardia	Fibroma odontogênico Tumor	Tumor odontogenico Fibroma
Melhoramento	Melhoramento Melhoramento genetico vegetal	Cálculo salivar gigantes em ducto de glândula submandibular	Calculos dos ductos salivares
Nutrição	Nutricao Nutricao vegetal	Hiperplasia	Hiperplasia gengival

(continua)

(conclusão)

Dificuldade em atribuir termos compostos			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
Análise	Analise foliar	Radiografia digital	Radiografia dentaria digital Radiografia dentaria digital direta USE Radiografia dentaria digital Radiografia digital dentaria USE Radiografia dentaria digital Radiografia digital USE Intensificacao de imagem radiografica
Praga Planta	Praga de planta	Carcinoma	Carcinoma de celulas escamosas
Distribuição Distribuição espacial	Distribuicao geografica	Adesivos	Adesivos dentinarios
Leprose Laranjeira	Leprose citrica	Resistência	Resistencia ao cisalhamento
Características químicas	Composicao quimica	Diagnóstico	Diagnostico bucal
Indução Brotação	Brotacao induzida	Micronúcleos	Testes para micronucleos
Taxa respiratória	Respiracao Taxa	Ionômero de vidro	Cimentos de ionomeros de vidro
		Neoplasias Glândulas salivares	Neoplasias das Glandulas salivares
		Regeneração	Regeneracao ossea
		Estudo Estudos	Estudos transversais
		Prótese total	Protese total superior Protese total inferior
		Infecção	Infeccoes oportunistas
		Desmineralização	Desmineralizacao do dente

Fonte: Elaborado pela autora

No QUADRO 32, constatamos que o sistema possui dificuldade em atribuir termos compostos porque, em geral, os termos não são apresentados no artigo na forma como estão no vocabulário controlado. No artigo, o autor menciona inicialmente o termo composto e, a partir daí, utiliza apenas a sua primeira unidade lexical para lhe fazer referência.

Relacionada a esse comportamento dos autores dos artigos, verificamos uma distinção entre a composição dos termos compostos do vocabulário DeCS em relação ao ThesAgro. Os termos compostos do DeCS apresentam elevada coordenação entre seus componentes, o que dificulta a atribuição automática.

É necessário lembrar que o uso de termos compostos é uma característica comum nas áreas especializadas e que fragmentá-los pode desconstruir seu significado. Dentre as propostas metodológicas para identificação automática de termos compostos, Café (2003) investigou como são constituídos, constatando que, em sua maioria, constituem-se por uma base, argumento e/ou satélites — Café denominou esses constituintes como "Unidades Terminológicas Complexas".

Verifica-se também o desenvolvimento de sistemas para a identificação de sintagmas nominais por processos automáticos como uma alternativa para solucionar a dificuldade de atribuir termos compostos, uma vez que os sintagmas nominais são unidades compreendidas como a menor parte do discurso portadora de informação, e que, ao ser extraída do texto, mantém o seu significado (KURAMOTO, 2006).

Para Anderson & Pérez-Carballo (2001), a identificação de termos compostos também pode ser útil para identificar nomes próprios: nomes de pessoas, organizações, países, marcas utilizados em determinados tipos de pesquisa.

d) Diferença na apresentação entre os termos do artigo e do vocabulário controlado

Este fator ocorre quando o termo apresenta uso de sinais, símbolos e, também, quando se verificam sutis diferenças entre termos do artigo e do vocabulário controlado que impedem a atribuição de termos de indexação.

QUADRO 33 - Fator interveniente na aplicação do ThesAgro e DeCS (diferença na apresentação entre os termos do artigo e do vocabulário controlado)

Diferença na apresentação entre os termos do artigo e do vocabulário controlado			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
Porta-enxerto	Porta enxerto	Adenoma pleomórfico	Adenoma pleomorfo
Jambo-vermelho	Jambo	Retrobturação	Obturacao retrograda
Quebra de dormência	Quebra da dormencia	Mcrodureza	Dureza
Caractere agronômico Caractere	Caracteristicas agronômicas	Leucoplasia oral	Leucoplasia Bucal
Maracujazeiro-amarelo	Maracujazeiro USE Maracuja	Oncogênese	Oncogenes
Cercas-vivas	Cerca viva USE Planta para cerca viva	Cirurgião dentista	Odontologos
‘Ponkan’	Ponkan	Técnica da citologia	Tecnicas citologicas
Passiflora	Passifloracea	Câncer de boca	Cancer da boca USE Neoplasias bucais
Banana ‘Maçã’	Banana Maca	Exame clínico	Exames medicos
Vida de prateleira Vida útil pós-colheita	Vida-de-prateleira	Câncer da região de cabeça e pescoço	Cancer de cabeça e pescoco USE Neoplasias de cabeça e pescoco
Mosca-das-frutas	Mosca das frutas	Síndrome da ardência bucal	Sindrome da boca ardente
Morte-precoce	Morte precoce	Materiais restauradores estéticos	Materiais dentarios
K e Ca	Potassio e Calcio	Índice de cárie	Indice CPO
Goiabeira-serrana	Goiaba serrana	Auxiliares da odontologia	Auxiliares de odontologia
(Myrtaceae)	Myrtaceae	Líquen plano oral	Liquen plano bucal
		CPOD	Indice CPOD USE Indice CPO
		Fatores sociais e econômicos	Fatores socioeconomicos
		Gengivoestomatite herpética	Estomatite herpetica
		Câncer de laringe	Cancer da laringe USE Neoplasias laringeas
		Erosão dental	Erosao dentaria
		Frutas cítricas	Citrus
		Oclusão balanceada bilateral	Oclusao dentaria balanceada
		Materiais para modelagem	Materiais para moldagem odontologica
		Imuno-histoquímica	Imunohistoquimica USE Imunoistoquimica

Fonte: Elaborado pela autora

No QUADRO 33, verificamos como a diferença de padrões entre termos que possuem o mesmo conceito impede a atribuição de termos de indexação. Constatamos as seguintes situações: uso de preposição, uso de hífen, de aspas, de parênteses ou de outros sinais gráficos, uso de símbolos, uso de sinônimos ou quase sinônimos.

A norma ANSI/NISO Z39.19-2005 recomenda que o hífen e os parênteses sejam evitados na elaboração dos vocabulários controlados, salvo exceções em que a sua exclusão possa ocasionar ambiguidade. Os vocabulários controlados ThesAgro e DeCS atendem a recomendação da norma, mas esses caracteres são utilizados nos termos apresentados nos artigos científicos. Portanto, é a distinção entre o termo do vocabulário e o termo do artigo que impossibilita a atribuição de termos de indexação.

Brooks (1998) e Anderson & Pérez-Carballo (2001) mostram que o uso de sinais no texto deve ser identificado por sistemas, justamente porque a interferência ocorre sobre os aspectos semânticos dos textos, constatando que a falta de normalização ou a sua identificação pode comprometer a recuperação da informação.

Como solução ao problema ocasionado pelo uso de hífen, Anderson & Pérez-Carballo (2001) propõem a alternativa de apresentar todas as possíveis combinações das palavras e ao tratar do uso de parênteses considerá-los como parte das palavras.

Verificamos que as aspas simples e os parênteses foram utilizados para identificar variedades de plantas e nomes científicos, mas que não são utilizados de forma padronizada. Torna-se apenas um indicativo, mas não uma regra amplamente aplicável.

Anderson & Pérez-Carballo (2001) também apresentam outros elementos, tais como o uso de números, a identificação de palavras constituídas por apenas um caractere e a identificação de letras maiúsculas e minúsculas. A identificação de números pode ser importante, dependendo da área de conhecimento em que será aplicado o sistema; as palavras constituídas por um caractere podem ser significativas no contexto do assunto tratado no artigo e não devem constar na relação de palavras vazias; por sua vez, as letras maiúsculas e minúsculas podem indicar nomes próprios e permitir a identificação de nomes de pessoas, organizações, países, etc., assim como abreviaturas como “Dr.”, “Ms.”, “Prof.” indicam que depois existe um nome próprio (GIL LEIVA, 2008).

No QUADRO 33, identificamos o uso de siglas e símbolos. O excesso de siglas, abreviaturas e símbolos nos textos foi constatado na análise de laudos médicos em pesquisa de Ferneda, Galvão e Rocha (2010). Como alternativa, foram normalizadas as palavras por

meio de uma lista pré-definida constituída por siglas, símbolos e abreviaturas e a palavra normalizada correspondente.

Identificamos também o uso de preposições diferentes entre os termos de indexação. Como destacado por Café (2003) no estudo das UTCs, as preposições desempenham um papel preponderante na definição do significado dos termos compostos.

e) Dificuldade em atribuir conceitos implícitos

Alguns termos de indexação não são atribuídos porque os conceitos a que se referem estão representados implicitamente no artigo científico.

QUADRO 34 - Fator interveniente na aplicação do ThesAgro e DeCS (dificuldade em atribuir conceitos implícitos)

Dificuldade em atribuir conceitos implícitos			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
... armazenado em câmaras frias ...	Refrigeracao	... 3ª, 4ª e 5ª décadas de vida ... faixa etária ...	Distribuicao por idade
... reprodução ... sistema reprodutivo ... polinização ...	Reproducao vegetal	... osso ... ossos ... tumores...	Osso USE Osso e Ossos Tumores USE Neoplasias Neoplasias osseas
... temperatura ... ar ... umidade ... iluminação ...	Climatologia	... obstrução das vias aéreas superiores ...	Obstrucao nasal
... cultivares ... características culturais ... cultivo ... ciclo de produção ...	Pratica cultural	... doença ... patologia ... tecido esquelético ... distúrbio ósseo ...	Doenças osseas
... induzir a brotação ... estimular a brotação ...	Brotacao induzida	... doença da Iga ... origem auto-imune ... enfermidade	Doenças auto-imunes
... fertilização ... nitrogênio ...	Fertilizante nitrogenado	... programa ... ação educativa ... campanha ... saúde ...	Promocao da saude
... bactéria ... inseto ... infecção ... plantas ...	Praga de planta	... fibroma ossificante (indica um tipo de neoplasia óssea) ...	Neoplasias osseas Fibroma Ossificante
... comportamento de novas cultivares ...	Comportamento de variedade	... idosos ... pacientes idosos ... saúde periodontal ...	Odontologia geriatrica

(continua)

(conclusão)

Dificuldade em atribuir conceitos implícitos			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Artigo científico	**Vocabulário controlado**	**Artigo científico**	**Vocabulário controlado**
... características químicas ... características físicas ...	Propriedade físico-quimica	... pacientes geriátricos ... idosos ... doenças crônicas ... população idosa ... avaliação ... saúde bucal ...	Avaliacao geriatrica Odontologia geriatrica Assistencia odontologica para doentes cronicos
... enxerto ... produção de mudas ... enraizamento de estacas ... estaquia de ramos ...	Propagacao vegetativa	... faixa etária ... idade média ... idade ...	Fatores etarios
... danos mecânicos ... dano físico ... dano externo ...	Dano mecanico	... mulheres ... sexo ...	Fatores sexuais
... pós-colheita ... embalagem ...	Conservacao de alimento USE Preservaçao de alimento	... agentes clareadores ... clareamento dental ...	Clareamento de dente
... lepidópteros minadores ...	Lagarta minadora	... atendimento ... pacientes com necessidades especiais ...	Assistencia odontologica para pessoas portadoras de deficiencias
... efeito dos resíduos ... efeito tóxico ... efeito desses agrotóxicos ...	Efeito residual	... educação em saúde bucal ... técnica educativa ...	Educação em odontologia
... crescimento vegetativo ... desenvolvimento vegetativo ...	Propagacao vegetativa	... reabilitador ... reabilitadores protéticos ... reabilitação protética ...	Reabilitacao bucal
... variabilidade genética ... variabilidade intra-específica ...	Variacão genetica	... tintura fitoterápica ... substâncias naturais ... propriedades terapêuticas ...	Fitoterapia
... teores foliares ... avaliação nutricional ...	Analise foliar	... proteção do complexo dentina-polpa ... protetor do complexo dentina-polpa ... protetores pulpares ...	Capeamento da polpa dentaria
		... espessura do esmalte proximal à altura do ponto de contato ... correção de discrepâncias dentais ... valores do diâmetro mésio-distal ...	Odontometria

Fonte: Elaborado pela autora

No QUADRO 34, verificamos que os termos de indexação não foram atribuídos porque os conceitos a que se referem estão, de certa forma, implícitos no artigo.

Verificamos, por exemplo, em um artigo as expressões "*enxerto*", "*produção de mudas*", "*enraizamento de estacas*" e "*estaquia de ramos*", sinalizando que se trata do assunto "*propagação vegetativa*". Para um indexador humano, atribuir o termo "*propagação vegetativa*" é uma atividade simples; já para um sistema automático, torna-se difícil.

Uma análise interpretativa sobre os conceitos do artigo permite representá-los pelos termos indicados no vocabulário controlado. Esse processo exige compreensão e reflexão sobre os conceitos, tarefa simples para um indexador humano; ao contrário, um sistema automático precisa de uma rede bem estruturada de conceitos e da formalização do seu significado para poder inferir tal conhecimento.

f) Diferença semântica nos termos de indexação

Não existe correspondência semântica entre os termos do artigo e os do vocabulário controlado, resultando em atribuição de termos sem relação com o assunto tratado no artigo.

QUADRO 35 - Fator interveniente na aplicação do ThesAgro e DeCS (diferença semântica nos termos de indexação)

Diferença semântica nos termos de indexação					
Aplicação do ThesAgro no SISA			**Aplicação do DeCS no SISA**		
Artigo científico	**Vocabulário controlado**	**Termo atribuído por SISA**	**Artigo científico**	**Vocabulário controlado**	**Termo atribuído por SISA**
Pelo (preposição)	Pelo (pelagem)	Pelo	Cápsula (cápsula do cisto)	Capsula USE Frutas (Cápsula da planta)	Frutas
Pato Branco (município)	Pato (ave)	Pato			
Mato Grosso do Sul (Estado)	Mato (vegetação)	Mato			
Sr. Pinto (variedade de laranjeira)	Pinto (ave)	Pinto			
Pêra (variedade de laranjeira)	Pera (fruto da pereira)	Pêra			
Capão Bonito (município)	Bonito (espécie de peixe)	Bonito			
Rio Grande do Norte (Estado)	Rio (canal fluvial)	Rio			
Primor Amoreira (variedade de mangueira)	Amoreira USE Amora	Amora			
Rosa (variedade de mangueira)	Rosa (flor da roseira)	Rosa			

Fonte: Elaborado pela autora

O QUADRO 35 mostra a identificação de apenas um exemplo de diferença semântica entre os termos na aplicação do DeCs, ao contrário da aplicação do ThesAgro, que apresentou

vários exemplos que podem implicar na recuperação de artigos científicos irrelevantes para a pesquisa. A partir desses exemplos é possível verificar que, mesmo delimitada a área específica de fruticultura, podem ocorrer situações de ambiguidade quando o sistema atua sobre palavras.

g) Atribuição automática de termo geral e de termo específico

Não há distinção entre termo geral e específico, culminando na atribuição das duas formas de termos de indexação.

QUADRO 36 - Fator interveniente na aplicação do ThesAgro e DeCS (atribuição automática de termo geral e de termo específico)

Atribuição automática de termo geral e de termo específico			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Termo geral	**Termo específico**	**Termo geral**	**Termo específico**
Clima	Clima temperado	Diagnostico	Diagnostico diferencial
Universidade	Universidade Federal	Adenoma	Adenoma Pleomorfo
Eugenia	Eugenia Involucrata	Neoplasias	Neoplasias osseas
Areia	Areia fina	Diagnostico	Diagnostico por Imagem
Oleo	Oleo mineral	Neoplasias	Neoplasias Bucais
Disturbio	Disturbio fisiologico	Bacterias	Bacterias anaerobicas
Adubacao	Adubacao verde	Saude	Saude bucal
Casca	Casca de arroz	Fibroma	Fibroma Ossificante
Iluminacao	Iluminacao artificial	Artrite	Artrite reumatoide
Umidade	Umidade relativa	Resinas	Resinas Compostas
Planta	Planta hospedeira	Artrite	Artrite Reumatoide
Mercado	Mercado atacadista	Articulacao temporomandibular	Transtornos da articulacao temporomandibular
Laranja	Laranja pera	Tomografia	Tomografia Computadorizada por Raio X
Deficiencia	Deficiencia hdrica	Neoplasias Boca	Neoplasias bucais
Profundidade	Profundidade de semeadura	Biopsia	Biopsia por agulha

(continua)

(conclusão)

Atribuição automática de termo geral e de termo específico			
Aplicação do ThesAgro no SISA		**Aplicação do DeCS no SISA**	
Termo geral	**Termo específico**	**Termo geral**	**Termo específico**
acido	acido indolbutirico	Saliva	Saliva artificial
Latossolo	Latossolo amarelo	Resinas	Resinas acrilicas
acido	acido borico	Músculos	Musculos mastigatorios
Fisiologia	Fisiologia vegetal	Leucoplasia	Leucoplasia pilosa
Microscopia	Microscopia eletronica		

Fonte: Elaborado pela autora

O QUADRO 36 mostra os termos gerais e específicos atribuídos na indexação automática. Verificamos que os artigos se referem apenas ao termo específico, mas o sistema também reconhece os termos gerais devido ao fato de que os termos específicos (geralmente palavras compostas) ocorrem no artigo de forma fragmentada (isto é, o autor faz referência ao termo específico utilizando-se apenas da primeira palavra que o compõe, que corresponde ao que o sistema interpreta como termo geral).

Tal situação interferiu nos índices de consistência e conduz a elevado grau de exaustividade na recuperação da informação, em detrimento da precisão, uma vez que não se distinguem artigos que tratam de assuntos gerais e específicos.

h) Atribuição de termos relacionados à metodologia da pesquisa

Termos que se referem à metodologia empregada no desenvolvimento do estudo foram atribuídos porque apresentaram alta ocorrência nos artigos científicos.

QUADRO 37 - Fator interveniente na aplicação do ThesAgro e DeCS (atribuição de termos relacionados à metodologia da pesquisa)

Atribuição de termos relacionados à metodologia da pesquisa	
Aplicação do ThesAgro no SISA	**Aplicação do DeCS no SISA**
Trabalho	Estatística
Método estatístico	Lâminas
Análise	
Método	
Pesquisa	

(continua)

(conclusão)

Atribuição de termos relacionados à metodologia da pesquisa	
Aplicação do ThesAgro no SISA	**Aplicação do DeCS no SISA**
Areia fina	
Papel	
Estação experimental	
Escola	
Agar	
Tecnologia	
Mandioca	
Laboratório	
Vinagre	

Fonte: Elaborado pela autora

No DeCS, constatamos poucos exemplos de atribuição de termos relacionados à metodologia. Verificamos que o DeCS atribuiu termos que são muito recorrentes nos artigos da área de odontologia, como, por exemplo, os termos "*estudo de casos*", "*síndrome*", "*odontologia*, "*diagnóstico*". No ThesAgro, a atribuição de termos de indexação dessa natureza foi recorrente e teve influência sobre o baixo índice de consistência.

i) Relação de equivalência omitida

Alguns termos do vocabulário controlado não apresentam remissivas que permitiriam a atribuição de termos relevantes, a exemplo de outros termos da mesma natureza que são contemplados no vocabulário controlado.

QUADRO 38 - Fator interveniente na aplicação do ThesAgro (relação de equivalência omitida)

Relação de equivalência omitida	
Aplicação do ThesAgro no SISA	
Artigo científico	**Vocabulário controlado**
Lichieira	Lichia
Aceroleira	Acerola
Pequizeiro	Pequi
Pereira	Pera

Fonte: Elaborado pela autora

O fator “Relação de equivalência omitida” não foi identificado durante a aplicação do DeCS no SISA, mas apenas na aplicação do ThesAgro. A remissiva é um recurso importante dos vocabulários controlados. Nos sistemas de indexação automática, torna-se uma alternativa para contornar as diversas formas com que os assuntos são apresentados no texto e permite identificá-los e representá-los por apenas um termo autorizado do vocabulário controlado. Mas é necessário cautela em sua elaboração e em seu uso, porque um termo do texto pode não corresponder ao termo autorizado para o qual foi remetido. Por exemplo: muitos artigos de odontologia se referem a “profissionais” e o sistema atribui “Associação de profissionais”, por constar a remissiva “Profissionais USE Associacao de profissionais” no vocabulário controlado. Ou seja, por mais que o artigo aborde algum aspecto relativo aos profissionais da área de odontologia, o contexto do artigo indica que não se trata de um artigo sobre associação de profissionais.

No experimento com o uso do DeCS no SISA, os termos de indexação “*Respiração bucal*”, “*Doença*”, “*Infiltração*”, “*Contaminação*” e “*Movimentação dentária*” não foram atribuídos e não há motivo evidente. Supomos que a presença da letra “c” com sinal diacrítico “cedilha” (“ç”) interferiu na atribuição de termos de indexação[33]. A possibilidade dessa interferência foi constatada apenas durante os testes de aplicação do ThesAgro no SISA, o que nos permitiu reconfigurar o sistema.

7.1.2 Análise dos fatores intervenientes na recuperação da informação

A análise da recuperação da informação nas bases de dados foi realizada abordando-se as características de exaustividade e precisão. A interferência na exaustividade e na precisão esteve relacionada à omissão dos termos no processo de indexação, tornando impossível a recuperação de muitos artigos científicos relevantes para a busca. A precisão na busca também é prejudicada quando se recuperam artigos além dos considerados relevantes, mas, no caso das buscas realizadas, a precisão não foi tão prejudicada por este fator, porque não foram recuperados muitos dos que estabelecemos como mínimo necessário.

No QUADRO 39 é possível verificar a relação de fatores que interferiram na indexação e na recuperação em ambas as pesquisas:

[33] A letra “c” com sinal diacrítico cedilha (“ç”) não é reconhecida no processamento automático do sistema SISA, pois o sistema foi idealizado para o idioma espanhol, que não utiliza esse sinal.

QUADRO 39 - Interferência na indexação e recuperação da informação

Fatores	Aplicação do ThesAgro no SISA		Aplicação do DeCS no SISA	
	Indexação	**Recuperação da Informação**	**Indexação**	**Recuperação da Informação**
Atribuição automática de termo geral e de termo específico	Sim	Não	Sim	Não
Atribuição de termos relacionados à metodologia da pesquisa	Sim	Não	Sim	Não
Diferença na apresentação entre os termos do artigo e os do vocabulário controlado	Sim	Sim	Sim	Sim
Diferença semântica nos termos de indexação	Sim	Não	Sim	Não
Dificuldade em atribuir conceitos implícitos	Sim	Sim	Sim	Sim
Dificuldade em atribuir termos compostos	Sim	Sim	Sim	Sim
Frequência de ocorrência dos termos em apenas uma estrutura do documento	Sim	Sim	Sim	Sim
Relação de equivalência omitida	Sim	Sim	Não	Não
Termos no singular e no plural	Sim	Sim	Sim	Sim

Fonte: Elaborado pela autora

Verificamos que o fator “atribuição automática de termo geral e de termo específico” não interferiu na recuperação da informação. Ao analisar o assunto dos artigos, verificamos que, em realidade, o termo geral representa uma parte do termo específico, ou seja, o artigo trata apenas do assunto específico. Por exemplo: o vocabulário controlado possui o termo geral “Disturbio”, mas o artigo trata especificamente de “Distúrbio fisiológico”. O fator “dificuldade em atribuir termos compostos” explica que o sistema atua sobre a identificação de palavras isoladas, ao invés de identificar termos compostos. Daí se verifica por que o sistema atribui termos gerais, além dos específicos. O sistema de indexação automática não distingue o que seria geral ou mais específico.

Outro fator que não interferiu na recuperação da informação foi a “atribuição de termos relacionados à metodologia da pesquisa”, visto que, na elaboração das estratégias de busca, não contemplamos assuntos relacionados a esse fator. No entanto, é necessário frisar que esse fator pode interferir na busca, na medida em que artigos indexados sob esses termos de indexação não reflitam exatamente o assunto tratado no artigo e exijam um elevado esforço dos usuários durante a seleção dos resultados.

O fator “diferença semântica nos termos de indexação” também não interferiu, mas os exemplos identificados na análise da indexação ilustram os conflitos que pode gerar ao representar um assunto que não corresponda ao termo de indexação.

O fator “recuperação de artigos científicos além do estabelecido como relevante para a busca” foi identificado apenas na análise da recuperação da informação do experimento com aplicação do ThesAgro no SISA. A interferência está na precisão da informação, mas, na maioria dos casos analisados, o assunto é tratado no artigo como assunto secundário.

Expomos, em seguida, as possíveis variáveis que explicam a interferência dos fatores identificados na indexação e na recuperação da informação.

7.2 Análise das variáveis no processo de indexação automática

Os resultados da análise nos indicam a necessidade de ampliar essa apreciação associando as variáveis envolvidas no processo de indexação automática. Sob essa perspectiva, verificamos que os fatores intervenientes na indexação e na recuperação da informação são ocasionados por dificuldades relacionadas aos métodos de indexação automática estabelecido no sistema SISA e às características do vocabulário controlado.

7.2.1 Métodos de indexação automática

Os primeiros sistemas de indexação automática foram desenvolvidos com base em abordagens matemáticas, pautadas na extração de palavras. A palavra é considerada a unidade de representação da informação.

Todos os fatores intervenientes que identificamos são, de certa forma, consequência do critério de identificação de palavras como unidades de representação. O sistema SISA atua sobre a identificação de padrões de sequência de caracteres e não exatamente sobre os conceitos que as palavras representam. Por isso ocorrem casos em que conceitos específicos,

conceitos expressos por termos compostos e conceitos implícitos não são identificados pelo sistema.

A partir das décadas de 1970 e 1980, os estudos linguísticos foram impulsionados na área de indexação automática, buscando integrar analisadores linguísticos e matemáticos. Houve um avanço nessa área quando se passou de uma abordagem com enfoque na palavra como unidade de representação para um enfoque voltado à investigação de estruturas mais complexas, como os sintagmas nominais. Verifica-se a importância da identificação de conceitos na evidência de que o valor de uma análise automática se traduz na identificação do significado, das ideias expressas que podem estar explícitas e, da mesma forma, implícitas no texto dos documentos.

Ainda que analisadores morfológicos e sintáticos possam realizar a análise de estruturas linguísticas, a análise semântica não depende apenas dessas análises: exige, ainda, uma base de conhecimento. Kuramoto (2002) explica que existe uma grande diferença entre a indexação automática e a indexação pelo indexador humano, pois, na indexação realizada pelo ser humano, utiliza-se a base de conhecimentos da pessoa, do especialista, assim como as técnicas de análises de assunto e outras ferramentas (tesauros, vocabulários controlados, lista de termos). Por outro lado, na indexação automática a máquina não possui essa base de conhecimentos nem utiliza qualquer técnica de análise de assuntos, mas, tão somente, a extração de palavras isoladas dos documentos.

Os resultados da pesquisa mostram que a integração de analisadores morfológicos e sintáticos ao sistema SISA pode contribuir para melhorar os resultados de indexação. A lematização pode contribuir para solucionar os problemas relacionados ao fator “termos no singular e no plural”, ao normalizar as distinções de flexão de número, e ao fator “frequência de ocorrência dos termos em apenas uma estrutura do documento”, por constatar que muitos termos se apresentam de formas diferentes em diferentes estruturas do artigo.

Do mesmo modo, a *tokenização* é um recurso que pode contribuir para definir exatamente quais são as marcas (sinais, pontuações, hífens, parênteses, aspas, apóstrofo) que devem ser consideradas parte integrante dos termos de indexação. Exemplos foram apresentados no fator “diferença na apresentação entre os termos do artigo e os termos do vocabulário controlado”.

Embora os recursos de análise morfológica e sintática — tais como: lematização, *tokenização*, identificação de categorias gramaticais, eliminação de palavras vazias — ofereçam algumas soluções para indexação automática, existe uma grande dificuldade para

generalizar todas as situações que ocorrem na linguagem, pois cada idioma possui suas particularidades, seja quanto às regras gramaticais, seja quanto aos sistemas de escrita (o chinês e o árabe, por exemplo, possuem sistemas de escrita que não adotam o alfabeto latino).

Desse modo, constata-se que a indexação automática possui limitações e, portanto, é mais prudente direcionar a aplicação de sistemas de indexação automática a casos particulares que permitam definir melhor a escolha dos métodos de indexação.

Os métodos fundamentados nas ideias de Zipf e Luhn, ou seja, na frequência de ocorrência de palavras, foram importantes para o desenvolvimento dos primeiros sistemas de indexação automática e continuam a ser a base elementar dos sistemas atuais.

O critério de frequência no sistema SISA se estabelece a partir da combinação de frequências nas estruturas "título", "resumo" e "texto" do artigo. Nem sempre a frequência de um termo relevante ocorre em combinação de estruturas. É muito comum ocorrer apenas no texto e não alcançar a frequência requerida pelo SISA para poder atribuir o termo de indexação. Nesse sentido, outros critérios de ponderação dos termos para indexação poderiam ser integrados para torná-lo mais flexível.

Embora haja esforços para o desenvolvimento da indexação automática, os métodos têm-se limitado à análise de linguagem textual (ANDERSON & PÉREZ-CARBALLO, 2001). Atualmente, existe a facilidade de produção de recursos informacionais multimídias, mas não precisamos ir tão longe para afirmar a necessidade de que os sistemas identifiquem elementos além do texto, tais como fórmulas, números, imagens, gráficos, tabelas, legendas, etc., que podem ser encontrados em um simples documento.

Os artigos científicos de odontologia selecionados para o experimento com o uso do DeCs no SISA apresentavam muitas imagens, que foram desconsideradas no momento da conversão dos artigos para o formato TXT. A linguagem textual que acompanha a imagem pode até mesmo ser descontextualizada quando se desvincula da imagem. É importante desenvolver pesquisas sobre metodologias para análise automática de elementos além do texto, afinal, o assunto dos documentos é expresso pelo conjunto de informações manifestadas em diversas formas.

Verificamos que, no SISA, o formato TXT não reconhece elementos como imagens, tabelas, figuras, etc. A maior parte dos artigos científicos estão disponíveis em formato PDF, DOC e XML, entre outros formatos, o que indica a necessidade de que os sistemas de indexação permitam a inclusão de diversos formatos de documentos, tanto para preservar os elementos do documento como para facilitar o trabalho de configuração do sistema.

Constatamos que cada conceito relacionado à indexação — indexação assistida por computador, indexação semiautomática e indexação automática — reflete processos distintos de análise e representação da informação, o que pode também proporcionar resultados de indexação diferentes. Em um contexto mais amplo, podemos afirmar que cada conceito reflete a concepção de distintas políticas de indexação e as posturas profissionais do indexador. Na indexação assistida por computador, o indexador humano tem o papel de analisar o conteúdo, utilizando o sistema apenas para inserir os termos de indexação. Na indexação semiautomática, o papel do indexador humano é avaliar os termos que foram propostos pelo sistema. Na indexação automática, todo o processo é realizado pelo sistema. No entanto, o papel do indexador humano será projetar, desenvolver, aperfeiçoar e atualizar o sistema de indexação automática de tal forma que os resultados gerados pelo sistema sejam confiáveis. Quando se pensa em uma indexação semiautomática deve-se refletir também sobre os critérios e sobre os requisitos empregados para realizar essa avaliação.

A exigência para desenvolver métodos mais complexos atualmente é motivada pela necessidade que se verifica diante do contexto caótico de disponibilização de informações, como, por exemplo, o ambiente *Web*, e pelo contexto favorável ao desenvolvimento de ferramentas oferecido pelos avanços tecnológicos, que, esperamos, possam um dia auxiliar efetivamente no processo de indexação e recuperação da informação.

7.2.2 Vocabulário controlado na indexação automática

Os vocabulários controlados são instrumentos fundamentais para garantir a consistência na indexação. Passam a ser integrados nos sistemas de indexação automática para auxiliar no controle terminológico.

Os vocabulários controlados foram originalmente concebidos para o processo de indexação manual, ou seja, foram construídos para que os profissionais, por meio de uma análise reflexiva, pudessem atribuir o melhor termo para representar o assunto de que trata o documento.

Os primeiros vocabulários controlados, os cabeçalhos de assuntos, remontam ao final do século XIX, época em que os computadores ainda não eram utilizados. Após o surgimento dos tesauros, os cabeçalhos de assuntos, concebidos para uso nos catálogos alfabéticos de assuntos das bibliotecas, adotaram algumas características daqueles, justamente para poder dispor de uma estrutura mais flexível em termos de estabelecimento de relações semânticas e organização sistemática.

O DeCS foi desenvolvido a partir do MeSH (*Medical Subject Headings) da United States National Library of Medicine* (NLM). O MeSH, publicado em 1960, é um tesauro formado por uma lista de descritores representados também na forma de cabeçalhos de assuntos na área de Ciências da Saúde para a indexação e a recuperação de artigos de periódicos publicados nos Estados Unidos e em mais de 70 países, disponibilizados na base de dados MEDLINE (BOCATO, 2005).

O DeCS é um vocabulário controlado trilíngue criado e mantido pela BIREME para servir como linguagem única na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais, assim como para ser usado na pesquisa e na recuperação de assuntos da literatura científica nas fontes de informação disponíveis na Biblioteca Virtual em Saúde (BVS) (CENTRO LATINO-AMERICANO E DO CARIBE...).

Por sua vez, o ThesAgro foi concebido de acordo com as diretrizes da UNESCO, das normas do "*Principles directeurs pour Létablissement et le développement the thesaurus monolíngues*" (SC/WS/555, Paris, 1973), tendo sido lançada, a sua primeira versão, em 1979. É um tesauro especializado na literatura agrícola, aplicado à indexação e à recuperação de documentos, e foi desenvolvido pela BINAGRI (BIBLIOTECA NACIONAL DE AGRICULTURA...).

Verificamos que o DeCS apresenta elevada coordenação dos termos de indexação, mas alguns termos compostos ainda assim foram atribuídos. Isso pode ser explicado porque a elaboração do vocabulário controlado apoiou-se no princípio da garantia literária, característica comum na elaboração dos cabeçalhos de assuntos. O princípio da garantia literária indica que os assuntos sejam definidos em função de como são apresentados na literatura da área e de como serão buscados pelos usuários.

No entanto, observamos que, de um modo geral, a característica de coordenação do vocabulário DeCS também impediu que o sistema atribuísse uma quantidade elevada de termos de indexação para cada artigo. Isso pode explicar a diferença de termos atribuídos na indexação com o uso do DeCS e com o uso do ThesAgro.

Na aplicação do DeCS houve uma variação de 1 a 11 descritores e uma média de 5 a 6 descritores na indexação no SISA, com variação de 2 a 13 e com média de 4 a 5 descritores na indexação pela BIREME. Já na aplicação do ThesAgro obteve-se uma variação de 4 a 25 descritores e uma média de 14 descritores atribuídos pelo SISA, com variação de 3 a 14 descritores e média de 6 a 7 descritores atribuídos pela BINAGRI. A quantidade de

descritores é considerada no cálculo de consistência na indexação, o que explica, portanto, o índice mais elevado no experimento com o DeCS no SISA.

Verificamos que o ThesAgro é caracterizado por termos constituídos por apenas uma unidade lexical e por termos compostos, mas em menor medida do que o DeCS. Desse modo, no experimento com o ThesAgro foi atribuída uma elevada quantidade de termos de indexação porque a possibilidade de haver compatibilidade entre os termos simples que se encontram no artigo e no vocabulário controlado são maiores.

A questão que permanece quando se realiza um processo de cotejamento entre o vocabulário controlado e os artigos é se o conceito apresentado pelo termo do artigo científico corresponde ao conceito definido pelo termo do vocabulário controlado. Esse questionamento também se aplica à indexação humana, porque, de fato, os vocabulários controlados são limitados, desatualizados e específicos demais. Quando se trabalha com palavras, isto é, com uma sequência de caracteres expressa entre determinadas marcas, a garantia da expressão semântica se torna questionável.

O uso de vocabulário controlado de uma área específica não minimiza os riscos de ocorrer ambiguidade, haja vista os casos em que identificamos diferença semântica entre os termos do artigo e os do vocabulário controlado na área de fruticultura. Além disso, verificamos que a delimitação do vocabulário controlado a uma área específica pode se tornar um empecilho à indexação de assuntos interdisciplinares.

Na aplicação de sistemas de indexação automática em áreas específicas do conhecimento, a análise das características que as definem — como terminologia, tipologias documentais (artigos científicos, relatórios, acórdãos, livros, legislação, laudos médicos, etc.) e sua relação com as estruturas textuais, comportamento da produção científica da área, redes de colaboração de pesquisadores, relação temáticas entre pesquisas, o que indica tendências de pesquisas, rede de citações, coocorrência de assuntos — pode contribuir para o estabelecimento de requisitos para o aperfeiçoamento do sistema.

As características de cada área do conhecimento precisam ser analisadas, para que esses métodos de indexação automática possam ser aplicados com mais confiabilidade a áreas que contemplam características adaptáveis a tais métodos.

Analisamos também diversos sistemas de indexação automática multilíngues. Com o acesso às bases de dados por meio das redes de computadores, torna-se imprescindível dispor de sistemas capazes de lidar com várias línguas. As investigações de tradutores automáticos avançaram e estão sendo integradas aos sistemas de indexação automática com uso de

vocabulários controlados multilíngues. É necessário lembrar que muitos documentos apresentam termos em outras línguas, dependendo da área de conhecimento de que tratam e, assim, considerar esses termos na indexação pode ser relevante. Desse modo, será necessário refletir sobre a aplicação e a integração de vários vocabulários controlados em um mesmo sistema de indexação.

A atualização dos vocabulários controlados é um grande desafio e uma característica fundamental para acompanhar o desenvolvimento científico das áreas do conhecimento. Na indexação automática, metodologias de atualização integradas aos processos de análise automática dos textos podem ser empregadas para agilizar a atualização e aproveitar os assuntos novos que são identificados na literatura da área do conhecimento.

Pouliquen, Steinberger e Ignat (2003) explicam que a simples extração de palavras do texto e o seu cotejamento com o vocabulário controlado não é suficiente no processo de atribuição de termos de indexação. Isso se deve à constatação de que a maior parte dos termos atribuídos aos documentos não estão explicitamente apresentados no texto.

No SISA, o cotejamento entre o artigo científico e o vocabulário controlado teria resultados favoráveis se o sistema realizasse um tratamento morfológico e sintático no início da análise automática para tratar as distintas formas com que os termos do artigo se apresentam. A partir de dados textuais normalizados, o sistema teria condições de atuar por processo de cotejamento. É claro que o tratamento linguístico a que nos referimos deve considerar os problemas envolvidos com recursos como lematização, tokenização, eliminação de palavras vazias, tal como foi exposto nos capítulos anteriores.

O SISA apresenta limitação quanto ao uso das potencialidades de um tesauro ao permitir a configuração do vocabulário controlado na ordem alfabética dos descritores e contemplar apenas as relações de equivalência.

Os vocabulários controlados também precisam ser adaptados com relações semânticas que os sistemas possam reconhecer, distinguindo entre conceitos gerais e específicos, conceitos associados e equivalentes.

7.3 Aspectos de adaptação de vocabulário controlado na indexação automática

Em síntese, constatamos que os aspectos apresentados a seguir merecem uma análise especial quando aplicados os vocabulários controlados no processo de indexação automática:

a) Método de identificação das unidades representativas da informação

É necessário compreender como o sistema atua sobre as unidades de representação da informação, se sobre palavras, sobre conceitos ou sobre o contexto a que os elementos textuais se referem. Esse aspecto é importante para se verificar em que medida o vocabulário controlado contempla as relações entre os conceitos, se possui estrutura que permite criar uma rede de conceitos, ou se os termos são constituídos com base no princípio da garantia literária para que o sistema consiga atuar sobre as palavras.

b) Aplicação prévia de tratamento linguístico

O tratamento linguístico permite que o vocabulário controlado atue sobre dados textuais normalizados quanto às distinções de natureza morfológica, sintática e semântica. Isso minimizaria os problemas apresentados nos experimentos que realizamos.

c) Relação entre o vocabulário controlado e os instrumentos de indexação automática

É necessário analisar a relação do vocabulário controlado com os outros instrumentos aplicados na indexação automática, como, por exemplo, a lista de palavras vazias, já que o tratamento preliminar de eliminação das palavras vazias pode impedir que o sistema atribua termos de indexação constituídos por alguma palavra considerada nessa lista.

d) Relações semânticas no vocabulário controlado

As relações de equivalência possuem uma função preponderante, pois ampliam a possibilidade de atribuir termos com o mesmo conceito, mas apresentados de formas distintas. Apesar de as áreas técnicas e científicas possuírem uma terminologia estabelecida constituída por termos técnicos, é muito comum o uso de sinônimos, de siglas e de abreviaturas que precisam ser, de alguma forma, apresentados explicitamente.

e) Características da área do conhecimento

As características da área do conhecimento podem fornecer indicações sobre a formação da terminologia da área quanto à constituição dos termos simples e compostos e dos conceitos objetivos e subjetivos, para que se possa verificar a forma com que serão tratados automaticamente.

Podem também auxiliar na análise da necessidade de cobertura entre a área específica do conhecimento e áreas correlatas, para contemplar a interdisciplinaridade de assuntos.

Algumas áreas produzem tipologias documentais constituídas por estruturas textuais específicas, que podem indicar mais precisamente a localização dos assuntos nos documentos e auxiliar a análise automática.

Da mesma forma, as características específicas da área podem indicar se é relevante para a análise integrar métodos de indexação automática que identifiquem elementos do documento como imagens, sons, esquemas, mapas, gráficos, fórmulas, quadros, tabelas, legendas, símbolos, etc.

f) Idioma do vocabulário controlado

Sistemas de indexação têm trabalhado com vocabulários controlados multilíngues por considerar que os textos dos documentos podem apresentar termos importantes para indexação em outros idiomas. Da mesma forma, vocabulários multilíngues são utilizados para apoiar a interoperabilidade entre bases de dados em idiomas diversos e minimizar os problemas de acesso à informação ocasionados por barreiras linguísticas.

g) Atualização do vocabulário controlado

É necessário estabelecer uma metodologia de atualização do vocabulário controlado para que, à medida em que novos conceitos surjam em determinada área do conhecimento, esses conceitos sejam incorporados ao vocabulário, impedindo a omissão de termos de indexação por ausência de termos no vocabulário.

h) Uso conjunto de vários vocabulários

É necessário integrar diversos vocabulários, uma vez que sistemas multilíngues precisam relacionar o vocabulário de cada idioma, assim como um sistema pode integrar vocabulários específicos para indicar nomes de organizações, eventos, etc. Para o tratamento linguístico, pode ser interessante incorporar um vocabulário constituído de radicais de palavras e, a exemplo das listas de palavras vazias, outro com termos muito comuns em determinada área de conhecimento para serem eliminados do texto.

Esses resultados nos permitem concluir que a complexidade envolvida na adaptação de um vocabulário controlado no processo de indexação automática exige o envolvimento de uma equipe multidisciplinar formada por profissionais de diversas áreas do conhecimento, tais

como Terminologia, Linguística, Estatística, Informática, Linguística Computacional e Ciência da Informação, para efetivamente poder contribuir na busca de soluções viáveis.

Sendo assim, apresentamos no próximo capítulo algumas considerações finais da pesquisa.

CONSIDERAÇÕES FINAIS

Esta pesquisa foi concebida a partir da constatação de que a complexidade envolvida no processo de indexação automática exige um olhar mais atento sobre as interferências causadas pela aplicação de vocabulários controlados na indexação automática.

Nesse contexto, investigamos a atuação dos vocabulários controlados no processo de indexação automática a partir da análise dos resultados da aplicação do vocabulário controlado ThesAgro no sistema SISA em artigos de periódicos científicos.

Tal proposta objetivou contribuir com o desenvolvimento da indexação automática ao levantar subsídios que nos permitiram identificar aspectos a respeito da adaptação de vocabulários controlados para aplicação no processo de indexação automática.

Nesse sentido, a partir da evolução dos instrumentos de indexação foi possível identificar e analisar as principais características que definem e distinguem os tipos de vocabulários controlados, constatando-se que a concepção dos cabeçalhos de assuntos é diferente da dos tesauros, embora se prestem ao mesmo objetivo, qual seja, o de auxiliar na organização e na recuperação da informação. Os cabeçalhos de assuntos foram concebidos para uso em catálogos de bibliotecas e os tesauros para uso em sistemas de recuperação da informação de centros especializados e, portanto, apresentam algumas distinções em suas características.

Embora essas características não tenham-se manifestado explicitamente nos resultados dos experimentos, podemos verificar uma sutil diferença entre a composição dos termos dos vocabulários controlados DeCS e ThesAgro, que interferiu na atribuição de termos de indexação no processo automático.

Examinamos os métodos adotados pelos sistemas de indexação automática e as alternativas propostas por pesquisas que buscam solucionar algumas dificuldades enfrentadas na indexação automática. Entre as propostas metodológicas, verificamos que, atualmente, muitos sistemas integram abordagens linguísticas: análise morfológica, sintática e semântica — esta ainda incipiente, com o emprego de instrumentos como os vocabulários controlados e as ontologias. Tal análise se prestou a verificar as possibilidades oferecidas pelas propostas metodológicas e pelos sistemas de indexação automática, no que se refere aos aspectos relacionados à aplicação de vocabulários controlados.

Dessa forma, foi desenvolvido o experimento com objetivo de aplicar e analisar o vocabulário controlado ThesAgro no processo de indexação automática do SISA, em análise comparativa com a indexação manual realizada pela BINAGRI. Da análise, foi possível constatar que os problemas na indexação automática do SISA estão relacionados a fatores de natureza linguística, *i.e.*, de tratamento morfológico, sintático e semântico, bem como relacionados à natureza metodológica do sistema e à aplicação do vocabulário controlado por simples processo de cotejamento entre os termos dos documentos e os do vocabulário.

A partir da identificação dos fatores que interferiram no experimento, foi possível analisar os fatores intervenientes na atuação dos vocabulários controlados ThesAgro e DeCS na indexação automática do SISA e na recuperação da informação em base de dados. Constatou-se que os fatores identificados na pesquisa com o DeCS são comprovados por experimento com o ThesAgro, apresentando apenas algumas variáveis diferentes.

Nesse sentido, sob uma perspectiva mais ampla dos problemas enfrentados na indexação automática, tivemos a comprovação de que o vocabulário controlado atua como instrumento que condiciona os resultados de indexação automática e, portanto, é extremamente importante compreendê-lo dentro de um processo em que estão envolvidos aspectos relacionados aos métodos de identificação das unidades representativas da informação, à aplicação de tratamento linguístico, à relação entre o vocabulário controlado e os instrumentos de indexação automática, às relações semânticas no vocabulário controlado, às características da área de conhecimento, ao idioma do vocabulário controlado, à atualização do vocabulário controlado e à possibilidade de uso conjunto de vários vocabulários controlados.

A análise das limitações verificadas na indexação automática e os resultados desta pesquisa permitem constatar que existe um grande desafio para efetivamente conseguir desenvolver sistemas que possam aproximar-se a capacidade de análise humana. Com certeza, poderão auxiliar na atividade de análise e serão aperfeiçoados à medida que o avanço tecnológico permitir.

No entanto, a indexação é uma atividade em que estão envolvidas diversas variáveis, que nem mesmo podem ser controladas por indexadores humanos, pois se trata de uma atividade que abrange a percepção da necessidade do outro.

Nesse contexto, podemos nos questionar se vale a pena automatizar para diminuir custos, bem como se é possível desenvolver sistemas capazes de realizar uma indexação orientada ao conteúdo como à demanda.

Muitas propostas de sistemas de indexação automática têm defendido a avaliação posterior dos termos de indexação por profissionais, percebendo que os resultados ainda não são totalmente confiáveis. Por isso, é importante desenvolver metodologias adequadas para avaliar o processo de indexação automática, bem como os resultados de indexação proporcionados por sistemas.

É necessário frisar que a indexação por humanos também pode apresentar algumas limitações, assim como a apresentada por indexação automática, se não houver a preocupação com a definição de uma política de indexação. Os critérios de avaliação por profissionais e o próprio desenvolvimento do sistema de indexação somente poderão ser estabelecidos de modo eficaz se houver a definição de uma política de indexação.

Constata-se que é preciso, antes de tudo, analisar as condições exigidas para possibilitar uma indexação de qualidade. Isso significa que será possível aplicar sistemas de indexação automática em algumas atividades, áreas, serviços, a tipos de documentos e, em outras, ainda será necessário contar com a capacidade de análise e reflexão, com o conhecimento prévio, a formação profissional e a experiência do indexador humano.

Portanto, esta pesquisa contribui com o desenvolvimento da área de indexação automática ao apontar os problemas que podem surgir e ao apresentar algumas das medidas investigadas para solucioná-los. Além disso, comprova a interferência de alguns fatores no processo de indexação automática e, por conseguinte, na recuperação da informação, levantando alguns aspectos relacionados ao uso de vocabulários controlados, em consonância com os métodos de indexação automática que merecem uma análise mais profunda.

Nesse sentido, consideramos que esta pesquisa é apenas uma “gota no oceano” dentre a vasta possibilidade de investigações que precisam ser desenvolvidas, em uma área ainda carente de reflexões teóricas e metodológicas.

Sendo assim, sugerimos que outros estudos possam se aprofundar sobre aspectos de relacionamentos semânticos em vocabulários controlados e em instrumentos como as ontologias aplicadas em indexação automática. Da mesma forma, acreditamos ser de grande valia realizar uma investigação sobre como as características de domínios específicos podem oferecer indícios que contribuam para a adaptação de sistemas de indexação e, dessa forma, permitam a aplicação de métodos mais confiáveis.

REFERÊNCIAS

ABAD GARCÍA, M. F. *Evaluación de la calidad de los sistemas de información*. Madrid: Sintesis, 2005. 202 p.

AITCHISON, J.; GILCHRIST, A. *Manual para construção de tesauros*. Rio de Janeiro: BNG/Brasilart, 1979. 142 p.

ALMEIDA, M. B. Roteiro para construção de uma ontologia bibliográfica através de ferramenta automatizada. *Perspect. Ciênc. Inf.*, Belo Horizonte, v. 8, n. 2, p. 164-179, jul./dez. 2003. Disponível em: < http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/40/176>. Acesso em: 27 de fev. 2011.

ANDERSON, J. D. & PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts of retrieval. Part II: machine indexing, and the allocation of human versus machine effort. *Information Processing and Management*, v. 37, n. 2, p. 255-277, mar. 2001.

BERNERS-LEE, T.; LASSILA, O.; HENDLER, J. The Semantic Web. Scientific America, maio de 2001. Disponível em: <http://www.sciam.com/article.cfm?id=the- semantic-web>. Acesso em 03/12/2008.

BIBLIOTECA NACIONAL DE AGRICULTURA. Disponível em: < http://www.agricultura.gov.br/portal/page?_pageid=33,958886&_dad=portal&_schema=PORTAL>. Acesso em: 30 de nov. 2010.

BOCCATO, V. R. C. *Avaliação de linguagem documentária em Fonoaudiologia na perspectiva do usuário*: estudo de observação da recuperação da informação com protocolo verbal. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2005.

BORST, W. N. *Construction of engineering ontologies for knowledge sharingand reuse.* 1997. Tese (Doutorado). Disponível em: <http:// www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf>. Acesso em: 25 jan. 2011.

BRÄSCHER, M. A ambigüidade na recuperação da informação. *DataGramaZero:* Revista de Ciência da Informação, v. 3, n. 1, fev. 2002. Disponível em: < http://www.dgz.org.br/fev02/Art_05.htm>. Acesso em: 04 de out. 2010.

BROOKS, T. A. Ortography as a fundamental impediment to online information retrieval. *Journal of the American Society for Information Science*, v. 49, n. 8, p. 731-41, 1998.

BUSH, V. As we may think. The Atlantic Monthly, Boston, MA, 1945. Disponível em: <http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush.shtml>. Acesso em: 22 de out. 2009.

CAFÉ, L. Contribuições da Gramática Funcional da delimitação de segmentos descritores de informação. In: RODRIGUES, G. M.; LOPES, I. L. (Org.). *Organização e representação do conhecimento na perspectiva da ciência da informação.* Brasília: Thesaurus, 2003, p. 118-140. (Estudos Avançados em Ciência da Informação, v.2)

CÂMARA JÚNIOR, A. T. da. *Indexação automática de acórdãos por meio de processamento de linguagem natural*. 2007. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília. Disponível em: <http://bdtd.bce.unb.br/tedesimplificado/tde_busca/arquivo.php?codArquivo=2403>. Acesso em: 20/06/2008.

CAMPOS, M. L. A. *Linguagens documentárias*: teorias que fundamentam sua elaboração. Niterói, RJ: EDUFF, 2001. 133 p.

CAMPOS, M. L. de A. *et al*. Estudo comparativo de softwares de construção de tesauros. *Perspect. ciênc. inf*., Belo Horizonte, v. 11 n. 1, p. 68-81, jan./abr. 2006. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/446/257>. Acesso em: 04 de dez. 2010.

CAMPOS, M. L.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. *Perspect. ciênc. inf*., v. 11, n. 3, p. 348-359, set./dez. 2006.

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

CENTRO LATINO-AMERICANO E DO CARIBE DE INFORMAÇÃO EM CIÊNCIAS DA SAÚDE. Bireme: fundamentos, missão, objetivos e funções. 2004. Disponível em: < http://www.bireme.br/local/Site/bireme/homepage.htm>. Acesso em: 20/10/2007.

CESARINO, M. A.; PINTO, M. C. M. F. Cabeçalhos de assunto como linguagem de indexação. *Revista da Escola de Biblioteconomia UFMG*, Belo Horizonte, v. 7, n. 2, p. 268-288, set. 1978.

CHAUMIER, J. *Analisis y lenguajes documentales*: el tratamiento lingüístico de la información documental. Barcelona: Ed. Mitre, 1986.

CHUNG, Y.; POTTENGER, W. M.; SCHATZ, B. R. Automatic subject indexing using an associative neural network. In: THIRD ACM CONFERENCE ON DIGITAL LIBRARIES, 1998, Pittsburgh, Pennsylvania, United States. *Proceedings...* Pittsburgh, 1998. p. 59-68. Disponível em: <[doi>10.1145/276675.276682]>. Acesso em: 16 de abr. 2010.

CINTRA, A. M. M. *et al. Para entender as linguagens documentárias*. 2. ed. rev. atual. São Paulo: Polis, 2002. 96 p.

CRAVEN, T. C. Linked phrase indexing. *Information Processing & Management*, v. 14, n. 6, p. 469-76, 1978.

CRAVEN, T. C. NEPHIS: a Nested-Phrase Indexing System. *Journal of the American Society for Information Science*, v. 28, n. 2, p. 107-14, mar. 1977. DOI: 10.1002/asi.4630280208.

CURRÁS, E. *Ontologias, taxonomias e tesauros em teoria de sistemas e sistemáticas*. Trad. Jaime Robredo. Brasília: Thesaurus, 2010.

CURRÁS, E. *Ontologias, taxonomias y tesauros*: manual de construccíon y uso. 3. ed. Gijón: Trea, 2005. 337 p.

DODEBEY, V. L. D. *Tesauro*: linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002. 119 p.

FERNEDA, E. Aplicando Algoritmos Genéticos na Recuperação de Informação. *DataGramaZero*: Revista de Ciência da Informação v. 10, n. 1, fev. 2009. Disponível em: < http://www.dgz.org.br/fev09/Art_04.htm>. Acesso em: 20 de mar. 2010.

FERNEDA, E. *Recuperação de informação*: estudo sobre a contribuição da Ciência da Computação para a Ciência da Informação. São Paulo, 2003. 147p. Tese (doutorado em Ciência da Informação). Escola de Comunicação e Artes, Universidade de São Paulo.

FERNEDA, E; GALVÃO, M. C. B.; ROCHA, J. E. S. Um método de indexação automática de documentos: aplicação em laudos de exames radiológicos. In: ENANCIB, 11, out. 2010, Rio de Janeiro, RJ. *Anais*...Rio de Janeiro: [s.n.], 2010. Disponível em: <http://congresso.ibict.br/index.php/enancib/xienancib/paper/view/491>. Acesso em: 30/10/2010.

FIORIN, J. L. *Introdução à Linguística*: I - Objetos teóricos. 2. ed. São Paulo: Contexto, 2003.

FOSKETT, A. C. *A abordagem temática da informação*. Trad. Antônio Agenor Briquet de Lemos. São Paulo: Polígono; Brasília: Ed.UnB, 1973.

FUJITA, M. S. L. *A análise documentária no tratamento da informação:* as operações e os aspectos conceituais interdisciplinares. Marília: Departamento de Ciência da Informação, 2003.

FUJITA, M. S. L. Avaliação da eficácia de recuperação do sistema de indexação Precis. *Ci. Inf.*, Brasília, v. 18, n. 2, p. 120-134, 1989.

FUJITA, M. S. L. *Linguagem documentária em odontologia:* uma aplicação do sistema de indexação PRECIS. São Paulo, 1992. Tese (doutorado em Ciências da Comunicação). Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 1992.

FUJITA, M. S. L.; RUBI, M. P. Um modelo de leitura documentária para a indexação de artigos científicos: princípios de elaboração e uso para formatação de indexadores. *DataGramaZero*: Revista Ciência da Informação, v. 7, n. 3, jun. 2006. Disponível em: < http://www.datagramazero.org.br/jun06/Art_04.htm>. Acesso em: 23 de dez. 2010.

FUNK, M. E.; REID. C. A. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, v. 71, p. 176-183, 1983.

GIL LEIVA, I, RUBI, M. P., FUJITA, M. S. L. Consistência na indexação em bibliotecas universitárias brasileiras. *Transinformação*, Campinas, v. 20, p. 233-54, 2008.

GIL LEIVA, I. Consistencia en la asignación de materias en bibliotecas públicas del Estado. *Boletín de la Asociación Andaluza de Bibliotecarios*, n. 63, p. 69-96, 2001. Disponível em: <http://webs.um.es/isgil>. Acesso em: 17 set. 2008.

GIL LEIVA, I. Consistencia en la indización de documentos entre indizadores noveles. *Anales de Documentación*, v. 5, p. 99-111, 2002. Disponível em: http://webs.um.es/isgil/. Acesso em: 26 de maio 2008.

GIL LEIVA, I. *La automatización de la indización de documentos*. Gijón: Trea, 1999. 221 p.

GIL LEIVA, I. *Manual de indización*: teoría y práctica. Gijón: Trea, 2008. 429 p. Sumário e prólogo da obra disponível em: <webs.um.es/isgil>. Acesso em: 23 de maio 2010.

GIL LEIVA, I.; RODRÍGUEZ MUÑOZ, J. V. Análisis de los descriptores de diferentes áreas de conocimiento. *Revista Española de Documentación Científica*, v. 20, n. 2, p. 150-160, 1997.

GIL URDICIAIN, B. *Manual de lenguajes documentales*. 2. ed. rev. e ampl. Gijón: Trea, 2004.

GIL, A. C. *Como elaborar projetos de pesquisa*. 3. ed. São Paulo: Atlas, 1996. 159 p.

GOMES, H. E.; MARINHO, M. T. Introdução ao estudo do cabeçalho de assunto. Disponível em: < http://www.conexaorio.com/biti/cabecalho/cab_ass.htm>. Acesso em: 26 de jan. 2011.

GOTTSCHALG-DUQUE, C. *SiRILiCO*: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia. 2005. Tese (Doutorado)- Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

GUIMARAES, J. A. C. A análise documentária no âmbito do tratamento temático da informação: elementos históricos e conceituais. In: RODRIGUES, G. M.; LOPES, I. L. (Org.). *Organização e representação do conhecimento na perspectiva da Ciência da Informação*. Brasília: Thesaurus, 2003, v. 2, p. 100-117.

GUIMARÃES, J. A. C. *Indexação em um contexto de novas tecnologias*. [S.L.:s.n.], 2000. 10 p. Texto didático.

IIVONEN, M.; KIVIMAKI, K. Common entibies and missing properties: similarities and differences in the indexing of concepts. *Knowledge Organization*, v. 25, n. 3, p. 90-102, 1998.

KIPP, M. E. I. Exploring Measures of Inter-Tagger Consistency. *SIG-CR Workshop Poster, Annual Meeting of the American Society for Information Science and Technology*, 2009, Vancouver, British Columbia, Canada. Disponível em: <http://eprints.rclis.org/17218/>. Acesso em: 11 de fev. 2011.

KOLAR, M., *et al*. Computer Aided Document Indexing System. *Journal of Computing and Information Technology*, v. 13, n. 4, p. 299-306, 2005.

KURAMOTO, H. Sintagmas nominais: uma nova abordagem no processo de indexação. In: NAVES, M. M. L.; KURAMOTO, H. (Orgs.). *Organização da informação:* princípios e tendências. Brasília: Briquet de Lemos, 2006.

KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. *DataGramaZero*: Revista de Ciência da Informação, v. 3, n. 1, fev. 2002. Disponível em: < http://www.dgz.org.br/fev02/Art_03.htm>. Acesso em: 08 de maio 2010.

LANCASTER, F. W. *El control del vocabulário en la recuperación de información*. 2. ed. Saragoza: Universidad de València; 2002. 281 p.

LANCASTER, F. W. *Indexação*: teoria e prática. 2. ed. rev. atual. Brasília, DF: Briquet de Lemos/Livros, 2004.

LANCASTER, F. W. *Vocabulary control for information retrieval*. 2. ed. Virginia : IRP, 1986.

LANCASTER, F.W. *Evaluation of the MEDLARS demand search service.* Washington, D.C. National Library of Medicine, 1968.

LARA, M. L. G. de. *Linguistica documentária*: seleção de conceitos. 2009. Tese (Livre-Docência) - Escola de Comunicação e Artes da Universidade de São Paulo, São Paulo, 2009.

LEININGER, K. Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science,* v. 32, n. 1, p. 4-8, 2000.

LEONARD, L. E. Inter-indexer consistency and retrieval effectiveness: measurement of relations. Champaign: University of Illinois, 1975 (PhD Thesis).

LIMA-MARQUES, M. *Ontologias:* da filosofia à representação do conhecimento. Brasília: Thesaurus, 2006. v. 1. 67 p.

LOPES, I. L. Uso de linguagens controlada e natural em bases de dados: revisão da literatura. *Ci. Inf.*, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

MARKEY, K. Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research,* v. 6, n. 2, p. 155-177, 1984.

MARTINHO, N. O. A dimensão teórica e metodológica da Catalogação de Assunto. 2010. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2010. Disponível em:< http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/martinho_no_me_mar.pdf>. Acesso em: mar. 2011.

MENDEZ RODRÍGUEZ, E. M.; MOREIRO GONZÁLEZ, J. A. Lenguaje natural e indización automatizada. *Ciencias de la Información*, v. 30, n. 3, p. 11-24, set., 1999.

MIDDLETON, M.R. A comparison of indexing consistency and coverage in the AEI, ERIC, and APAIS databases. *Behavioral & Social Sciences Librarian*, v. 3, n. 4, p. 33-43, 1984.

MONTEJO RAÉZ, A. Proyecto de indexado automático para documentos en el campo de la Física de Altas Energías. *Processamiento del Lenguaje Natural*, n. 27, p. 295-96, set. 2001.

MONTEJO RAÉZ, A. Towards conceptual indexing using automatic assignment of descriptors. In: 2 ND. INTERNATIONAL CONFERENCE ON ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB BASED SYSTEM. PERSONALIZATION TECHNIQUES IN ELETRONIC PUBLISHING ON THE WEB TRENDS AND PERSPECTIVES, 2002, Málaga, Spain. *Anais...*Málaga, Spain, maio 2002. p.115-20.

MOREIRO GONZÁLEZ, J. A. *El contenido de los documentos textuales:* su análisis y representación mediante el lenguaje natural. Gijón (Astúrias): Trea, 2004. 291 p.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de odontologia. *Inf. & Soc.: Est*., João Pessoa, v. 19, n. 2, p. 99-118, maio/ago. 2009.

NATIONAL INFORMATION STANDARDS ORGANIZATION (U.S.). *ANSI/NISO Z39.19-2005*: Guidelines for the construction, format and management of monolingual controlled vocabularies. Bethesda (USA): Ballot Period, abr./maio 2005. 172 p.

NESHAT, N.; HORRI, A. A study of subject indexing consistency between the National Library of Iran and Humanities Libraries in the Area of Iranian Studies. *Cataloging & Classification Quarterly*, v. 43, n. 1, p. 67-76, 2006.

NOVELLINO, M. S. F. Instrumentos e metodologias de representação da informação. *Inf.Inf*., Londrina, v. 1, n. 2, p. 37-45, jul./dez. 1996. Disponível em: < http://www.uel.br/revistas/uel/index.php/informacao/article/view/1603/1358>. Acesso em: 28 de mar. 2010.

NUBEL, R. *et al.* Bilingual Indexing for Information Retrieval with AUTINDEX. In: LREC *Proceedings*, Las Palmas, 2002. Disponível em:< http://www.iai.uni-sb.de/~bindex/IrecNuebel.pd>. Acesso em: 13 de jan. 2011.

OLIVEIRA, M. I. *Estudo do contexto de bibliotecas universitárias pelas abordagens de indexação e recuperação em domínios específicos*. Trabalho de Conclusão de Curso (Graduação em Biblioteconomia) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2009.

OTHERO, G. de A.; MENUZZI, S. de M. *Linguistica Computacional*: teoria & prática. São Paulo: Parábola Editorial, 2005.

PINTO MOLINA, M. *Analisis documental*: fundamentos y procedimientos. 2. ed. rev. aum. Madrid: EUDEMA, 1993.

POULIQUEN, B.; STEINBERGER, R.; IGNAT, C. Automatic annotation of multilingual text collections with a conceptual thesaurus. In: WORKSHOP AT EUROLAN'2003, jul./aug., 2003, Bucharest. *Anais*... Bucharest, 2003.

RAMALHO, R. A. S. *Desenvolvimento e utilização de ontologias em Bibliotecas Digitais: uma proposta de aplicação*. 2010. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: < http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/ramalho_ras_do_mar.pdf>. Acesso em: 24 de jan. 2011.

RIVIER, A. Construção de linguagens de indexação. *Revista da Escola de Biblioteconomia da UFMG*, v. 21, n. 1, p. 56-99, jan./jun.1992.

ROBREDO, J. *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. rev. e ampl. Brasília DF: Edição de autor, 2005. 410 p.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. *Ci. Inf.*, Brasília, v. 20, p. 130-6, jul.dez. 1991.

SAARTI, J. Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation,* v. 58, n. 1, p. 49-65, 2002.

SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. Nova York: McGraw-Hill, 1983.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. *Perspect. ciênc. inf.*, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SIEVERT, M. E.; ANDREWS, M. J. Indexing consistency in Information Science Abstracts. *Journal of the American Society for Information Science*, v. 42, n. 1, p. 1-6, 1991.

SILVA, M. R.; FUJITA, M. S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. *Transinformação*, Campinas, v. 16, n. 2, p. 133-161, maio/ago., 2004. Disponível em: http://revistas.puc-campinas.edu.br/transinfo/viewarticle.php?id=65>. Acesso em: 25/04/2007.

SILVESTER, J. P.; GENUARDI, M. T.; KLINGBIEL, P. H. Machine-aided indexing at NASA, *Information Processing & Management*, v. 30, n. 5, p. 631-45, 1994.

SOLER MONREAL, C.; GIL LEIVA, I. Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento: folksonomías, taxonomías y ontologías. *Revista Interamericana de Bibliotecología*, v. 33, n. 2, p. 361-77, jul/dez. 2010.

SOLER MONREAL, M. C. *Evaluacion de vocabulários controlados en la indizacion de documentos mediante índices de consistência entre indizadores*. 2009. Tese (Doutorado) – Universidad Politecnica de Valencia, Valência, 2009.

SOUZA, R. R. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais*. 2005. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais. Disponível em: < http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/RRSA-6GGGUF>. Acesso em: 20/06/2008.

STRAIOTO, A. C.; GUIMARÃES, J. A. C. A abordagem facetada no contexto da organização do conhecimento: elementos históricos. *Páginas a&b (arquivos & bibliotecas)*, Lisboa, n. 14, p. 109-136, 2004.

TONTA, Y. A Study of indexing consistency between Library of Congress and British Library Catalogers. *Library Resources and Technical Services*, v. 35, n. 2, p. 177-185, 1991.

TORRÊS, L. M. C. das. Sistematização da sintaxe de cabeçalhos de assunto. Disponível em: < http://www.conexaorio.com/biti/lecy/lecy.htm>. Acesso em: 26 de jan. 2011.

UNESCO. *Guidelines for the establishment and development of monolingual thesauri*. Paris, 1973. 37 p.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. *Ci. Inf.*, Brasília, v. 17, n. 1, p. 43-57, jan./jun., 1988.

VIZCAYA ALONSO, D. *Lenguajes documentários*. Rosario: Nuevo Paradigma, 1997.

WORLD INFORMATION SYSTEM FOR SCIENCE AND TECHNOLOGY. Princípios de indexação. *Revista da Escola de Biblioteconomia UFMG*, v. 10, n. 1, p. 83-94, 1981.

APÊNDICE A - Lista de Descritores (ThesAgro)

1. ABACA USE CANHAMO DE MANILLA
2. ABACATE
3. ABACATEIRO USE ABACATE
4. ABACAXI
5. ABACAXIZEIRO USE ABACAXI
6. ABASTECIMENTO
7. ABASTECIMENTO DE AGUA
8. ABATE
9. ABATEDOURO
10. ABELHA
11. ABELHA AFRICANA
12. ABELHA BRASILEIRA
13. ABELHA CARNICA
14. ABELHA CAUCASIANA
15. ABELHA EUROPEIA
16. ABELHA INDIGENA
17. ABELHA ITALIANA
18. ABELHA JATAI
19. ABELHA RAINHA
20. ABELMOSCHUS ESCULENTUS
21. ABIU
22. ABOBORA
23. ABOBORA CHEIROSA
24. ABOBORA D'AGUA USE CABACA
25. ABOBORA DE MOITA
26. ABOBORA HIBRIDA TETSUKABUTO USE ABOBORA JAPONESA
27. ABOBORA ITALIANA
28. ABOBORA JAPONESA
29. ABOBORA MENINA
30. ABOBORA MORANGA
31. ABOBORA RASTEIRA

.....

9530. ZINGIBERACEAE
9531. ZINIA
9532. ZINNIA ELEGANS
9533. ZIZIPHUS JOAZEIRO
9534. ZIZIPHUS JUJUBA
9535. ZIZIPHUS MAURITIANA
9536. ZOLLERNIA PARAENSIS
9537. ZOLLERNIA USE MOCITAIBA
9538. ZONA ARIDA
9539. ZONA BENTICA USE AMBIENTE BENTICO
9540. ZONA CLIMATICA
9541. ZONA INTERMARE
9542. ZONA LITORANEA USE LITORAL
9543. ZONA RURAL
9544. ZONA URBANA
9545. ZONEAMENTO AGRARIO
9546. ZONEAMENTO AGRICOLA
9547. ZONEAMENTO CLIMATICO
9548. ZONEAMENTO ECOLOGICO
9549. ZONEAMENTO FLORESTAL
9550. ZONEAMENTO PECUARIO
9551. ZOOGEOGRAFIA USE BIOGEOGRAFIA
9552. ZOOLOGIA
9553. ZOOLOGICO USE JARDIM ZOOLOGICO
9554. ZOONOSE
9555. ZOOPATOLOGIA USE DOENCA ANIMAL
9556. ZOOPHTHORA RADICANS
9557. ZOOPLANCTON
9558. ZOOTECNIA
9559. ZORNIA
9560. ZORNIA SPP
9561. ZOYSIA JAPONICA
9562. ZULIA ENTRERIANA
9563. ZYGOSACCHAROMYCES BAILLII

APÊNDICE B - Lista de palavras vazias

ABAIXO
ACASO
ACERCA
ACIMA
ADEMAIS
ADENTRO
ADEUS
ADIANTE
AFIRMA
AFIRMOU
AGORA
AÍ
AINDA
ALÉM
ALGUM
ALGUMA
ALGUMAS
ALGUNS
ALI
AMANHÃ
AMBAS
AMBOS
AMPLAMENTE
AMPLO
ANTE
ANTES
AONDE
AOS
APARECE
APARECEM
APARECER
APENAS
APESAR
APONTAR
APORTE
APÓS
APROXIMADAMENTE
AQUELA
AQUELAS
AQUELE
AQUELES
AQUI
AQUILO
ÁREA
AS
ÀS
ASSIM
ATÉ
ATRÁS
ATRAVÉS
BAIXO
BAIXOS
BASTANTE
BASTANTES
BILHÃO
BILHÕES
BOA
BOM
BONS
BREVE
BREVEMENTE
BUSCA
CÁ
CABE
CADA
CÂMBIO
CAPÍTULO
CAPÍTULOS
CARACTERE
CARACTERÍSTICA
CARACTERÍSTICAS
CATORZE
CAUSA
CEDO
CENTO
CENTOS
CERTA
CERTAMENTE
CERTAS
CERTEZA
CERTO
CERTOS
CIMA
CINCO
CINQÜENTA
COISA
COMIGO
COMO
COMPANHIA
COMPRIDO
CONDIÇÃO
CONFIÁVEIS
CONFIÁVEL
CONFORME
CONHECIDO
CONSEGUINTE
CONSEQÜÊNCIA
CONSIDERANDO
CONSIGO
CONTA
CONTÉM
CONTIGO
CONTRA
CONTRÁRIO
CONTUDO
CORRESPONDENTE
CORRESPONDENTES
CUJA
CUJAS
CUJO
CUJOS
CUMPRE
CUMPREM
CUSTA
DÃO
DAQUELA
DAQUELAS
DAQUELE
DAQUELES
DAR
DEBAIXO
DECIDIR
DÉCIMO
DECISÃO
DEMAIS
DEMASIADA
DEMASIADAS
DEMASIADO
DEMASIADOS
DENTRO
DEPOIS
DESDE
DESSA
DESSAS
DESSE
DESSES
DESTA
DESTAS
DESTE
DESTES
DETRÁS
DEVE
DEVEM
DEVERÁ
DEVERIA
DEZ
DEZENOVE
DEZESSEIS
DEZESSETE
DEZOITO
DIA
DIANTE
DIAS
DIREÇÃO
DISCREPÂNCIA
DISCREPÂNCIAS
DISPÕE
DISPÕEM
DISSE
DISSO
DISTANTE
DISTO
DITAS
DIZ
DIZEM
DIZER
DOBRO
DOIS
DONDE
DOZE
DUAS
DÚVIDA
É
ELA
ELAS
ELE
ELES
EMBAIXO
EMBORA
ENCIMA
ENCONTRA
ENCONTRAM
ENQUANTO
ENTANTO
ENTÃO
ENTRE
ENTRETANTO
ERA
ÉS
ESSA
ESSAS
ESSE
ESSES
ESTA
ESTÁ
ESTABELECE
ESTAMOS
ESTÃO
ESTARÁ
ESTAS
ESTAVA
ESTE
ESTES
ESTEVE
ESTIVE
ESTIVEMOS
ESTIVERAM
ESTIVESTE
ESTIVESTES
ESTOU
ESTUDADO
ESTUDADOS
ETC

EU
EXCETO
EXEMPLIFICAR
EXEMPLO
EXERCE
EXISTE
EXPLICAR
EXPRESSADO
EXPRESSAR
FAÇA
FACILITA
FACILITAR
FAÇO
FALTA
FARÁ
FATO
FAVOR
FAZ
FAZEIS
FAZEM
FAZEMOS
FAZER
FAZES
FAZIA
FEZ
FICOU
FIM
FINAL
FOI
FOMOS
FOR
FORA
FORAM
FORMA
FORMAS
FOSTE
FOSTES
FUI
GERAM
HÁ
HAVER
HAVIA
HOJE
INCLUSIVE
INICIAR
INÍCIO
INTELECTUALMENTE
IR
IRÁ
ISSO
ISTO
JÁ
JAMAIS
JUNTO
JUNTOS
LÁ
LADO
LADOS
LEVAR
LHE
LIGADO
LOGO
LONGE
LUGAR
MAIORIA
MAIORIAS
MALES
MANEIRA
MANEIRAS
MANHÃ
MAS
MAU
MÁXIMO
ME
MÉDIA
MEDIANTE
MELHOR
MENOS
MÊS
MESES
MESMA
MESMAS
MESMO
MESMOS
METADE
MEU
MEUS
MILHÃO
MILHÕES
MIM
MINHA
MINHAS
MODO
MODOS
MOMENTO
MOMENTOS
MOTIVO
MOTIVOS
MUITA
MUITAS
MUITO
MUITOS
NADA
NAQUELA
NAQUELAS
NAQUELE
NAQUELES
NAS
NEM
NENHUM
NENHUMA
NENHUMAS
NENHUNS
NESSA
NESSAS
NESSE
NESSES
NESTA
NESTAS
NESTE
NESTES
NINGUÉM
NOS
NÓS
NOSSAS
NOSSO
NOSSOS
NOVAMENTE
NOVAS
NOVE
NOVENTA
NOVOS
NUM
NUMA
NÚMERO
NUNCA
OBRIGADA
OBRIGADO
OBSTANTE
OBTÉM
OBTER
OFERECE
OFERECEM
OITAVA
OITAVO
OITENTA
OITO
ONDE
ONTEM
OPÇÃO
OPÇÕES
OS
OUTRA
OUTRAS
OUTRO
OUTROS
PARECE
PARES
PARTIR
PASSO
PASSOS
PEDIDO
PEDIDOS
PELA
PELAS
PERÍODO
PERMITE
PERMITIR
PERTO
PIOR
PIORES
PODE
PÔDE
PODEM
PODEMOS
PODERÁ
PÕE
PÕEM
POIS
PORQUE
PORQUÊ
POSSÍVEL
POSSIVELMENTE
POSSO
POSTERIOR
POUCA
POUCAS
POUCAS
POUCO
POUCOS
PRAZO
PREFERÊNCIA
PRETEXTO
PRIMEIRA
PRIMEIRAS
PRIMEIRO
PRIMEIROS
PRINCIPALMENTE
PRINCÍPIO
PRIORI
PROBLEMA
PROBLEMAS
PRONTO
PRÓPRIAS
PRÓPRIO
PRÓPRIOS
PRÓXIMA
PRÓXIMAS
PRÓXIMO
PRÓXIMOS
PUDERAM
QUAIS
QUAISQUER
QUAL
QUALQUER
QUANDO
QUANTA
QUANTAS
QUANTO
QUANTOS
QUÃO
QUARENTA
QUARTA
QUASE
QUATRO
QUE
QUÊ
QUEM

QUER
QUEREIS
QUEREM
QUERES
QUERO
QUINTA
QUINTO
QUINZE
REDOR
REGULAR
REPENTE
RESPEITO
RESTANTE
RESTANTES
SABE
SABEM
SABER
SALVO
SE
SEGUIDA
SEGUINDO
SEGUINTE
SEGUINTES
SEGUNDA
SEGUNDO
SEI
SEIS
SEJA
SEMPRE
SENDO
SER
SERÁ
SERÃO
SERIA
SESSENTA
SÉTIMA
SÉTIMO
SEU
SEUS
SEXTA
SEXTO
SIDO
SIM
SÓ
SOB
SOIS
SOMENTE
SOMOS
SÓS
SOU
SUA
SUAS
SUFICIENTEMENTE
TAIS
TAL
TALVEZ
TAMBÉM
TAMPOUCO
TANTA
TANTAS
TANTO
TANTOS
TÃO
TARDE
TE
TEL
TEM
TÊM
TEMOS
TENDES
TENHO
TENS
TENTARAM
TENTE
TENTEI
TER
TERÁ
TERCEIRA
TEU
TEUS
TEVE
TIVE
TIVEMOS
TIVERAM
TIVESTE
TIVESTES
TODA
TODAS
TODAVIA
TODO
TODOS
TRÁS
TRÊS
TREZE
TU
TUA
TUAS
TUDO
ÚLTIMO
ÚLTIMOS
UMA
UMAS
UNS
USA
USAR
VAI
VAIS
VÃO
VÁRIAS
VÁRIOS
VEM
VÊM
VENS
VER
VERDADE
VEZ
VEZES
VINDO
VINTE
VOCÊ
VOCÊS
VOS
VÓS
VOSSA
VOSSAS
VOSSO
VOSSOS

APÊNDICE C - Modelo de um artigo científico da área agrícola formatado segundo critérios do SISA[34]

#CTI#
BALANÇO DE CARBOIDRATOS EM GEMAS FLORAIS DE DOIS GENÓTIPOS DE PEREIRA SOB CONDIÇÃO DE INVERNO AMENO
#FTI#

#CR#
RESUMO - As pereiras européias e asiáticas, cultivadas sob condições de inverno ameno, como na região Sul do Brasil, apresentam problemas de adaptação. Durante o inverno, as oscilações térmicas e o baixo acúmulo de frio têm sido referidos por alguns autores como causas do abortamento de gemas florais. O objetivo deste trabalho foi determinar o balanço de carboidratos em tecidos de gemas florais de duas cultivares de pereiras: Kieffer (P. communis x P. pyrifolia) e Housui (P. pyrifolia). Os tecidos de gemas florais e da base de gemas foram coletados mensalmente, de fevereiro a setembro de 2002, de plantas de pomar da Embrapa Clima Temperado, Pelotas-RS, coordenadas 32°51' S e 52°21'O, localizado a 230 metros de altitude. O material vegetal foi analisado separadamente quanto às concentrações de açúcares solúveis (por cromatografia gasosa) e porcentagens de amido (por espectrofotometria). Em ambas as cultivares, observou-se que a base da gema é um importante local de reserva. Ocorreram significativos aumentos de açúcares solúveis nas gemas das duas cultivares na fase que antecede a brotação. Em setembro, os açúcares solúveis totais na matéria seca (MS), nas gemas florais da cv. Housui (38,33 mg g-1), foram menores do que os observados nos tecidos da cv. Kieffer (50,39 mg g-1), cultivar melhor adaptada às condições climáticas. O açúcar-álcool sorbitol, seguido da sacarose, foi o açúcar solúvel mais abundante nos tecidos das duas cultivares.
Termos para Indexação: Pyrus sp., açúcares solúveis, sorbitol, sacarose, amido, abortamento floral.
#FR#

#CTE#
INTRODUÇÃO
As plantas frutíferas de clima temperado apresentam o fenômeno da dormência. Nesse período, durante o inverno, ocorre a conversão do amido para açúcares solúveis, como substrato para a retomada de crescimento na primavera. O amido é o mais importante carboidrato de reserva nas plantas. No inverno, essas reservas amiláceas são parcialmente convertidas em açúcares solúveis dentro das partes aéreas e das raízes finas. Nas plantas frutíferas de clima temperado, as reservas são essencialmente utilizadas na primavera (Lacoite et al., 1993). A mobilização dos açúcares solúveis está diretamente ligada aos eventos climáticos, principalmente à temperatura, e tem grande importância nos estudos de adaptação de frutíferas de clima temperado. No Brasil, o estudo da mobilização dos carboidratos está sendo utilizado para compreender os problemas decorrentes da falta de frio hibernal, em frutíferas de clima temperado (Herter et al., 2001). A maioria das cultivares de pereiras produtoras de frutas de alta qualidade não tem boa adaptação às condições climáticas da região Sul do Brasil, principalmente, devido ao frio hibernal insuficiente para a satisfação da dormência (Petri et al., 2001). Disso, resulta brotação errática e deficiente, floração desuniforme com baixo número de flores, afetando negativamente a produtividade. Oscilações térmicas durante o inverno e baixo acúmulo de frio têm sido referidos como causa do abortamento de gemas florais (Herter et al., 2001). Nas cultivares de pereira mais exigente, o problema de abortamento ou necrose de primórdios atinge índices que, em algumas cultivares, inviabiliza a exploração econômica. Em geral, o abortamento é menos severo nas áreas mais frias, como São Joaquim-SC e Vacaria-RS.
Uma das hipóteses para a ocorrência do abortamento floral, é que os períodos com temperaturas relativamente altas durante o período de repouso das plantas (sem atividade fotossintética), poderiam provocar o aumento da taxa respiratória e exaurir as reservas de carboidratos em níveis insatisfatórios para suprir as necessidades das gemas florais para a retomada do crescimento e subseqüente floração, frutificação e emissão dos novos brotos da estação de crescimento (Gardin, 2002).
O presente trabalho teve como objetivo determinar os níveis dos carboidratos, no período de fevereiro a setembro, em gemas florais de duas cultivares de pereira, cultivadas em condições de inverno ameno.
MATERIAL E MÉTODOS
O experimento foi realizado na Embrapa Clima Temperado, em Pelotas-RS, no laboratório de Fisiologia Vegetal, em 2002. Utilizaram-se plantas de pereiras de duas cultivares, pertencentes à coleção instalada na Estação Experimental da Cascata, localizada nas coordenadas 32º 51' S e 52O 21' W e altitude de 230 m. A cultivar Kieffer (P. communis x P. pyrifolia) apresenta reduzido índice de abortamento, e a cultivar Housui (P. pyrifolia) apresenta problemas de adaptação e altos índices de abortamento quando crescidas em região de inverno ameno.
[...]
CONCLUSÕES
1. A base das gemas é um importante local de reserva em plantas de pereira das cvs. Kieffer e Housui. 2. Na fase que antecede a brotação, ocorreu significativo aumento de açúcares solúveis nas gemas das duas cultivares. 3. Em setembro, os açúcares solúveis totais na matéria seca, nas gemas florais da cv. Housui, foram menores do que os observados na cv. Kieffer, cultivar melhor adaptada às condições climáticas. 4. O sorbitol, seguido pela sacarose, foi o açúcar mais abundante em ambos os tecidos das duas cultivares de pereira.
#FTE#

[34] RODRIGUES, Alexandre Couto et al. Balanço de carboidratos em gemas florais de dois genótipos de pereira sob condição de inverno ameno. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 1-4. ISSN 0100-2945.

APÊNDICE D - Consistência entre a Indexação elaborada pela BINAGRI e por SISA

	Artigo Científico	Descritores (Indexação BINAGRI)	Descritores (Indexação SISA)	Consistência na Indexação
1	RODRIGUES, Alexandre Couto et al. Balanço de carboidratos em gemas florais de dois genótipos de pereira sob condição de inverno ameno. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 1-4. ISSN 0100-2945.	1. PERA 2. GEMA 3. GENÓTIPO 4. AMIDO 5. SACAROSE 6. CROMATOGRAFIA GASOSA 7. CARBOIDRATO 8. BROTAÇÃO 9. INVERNO	1.- INVERNO 2.- ACLIMATACAO 3.- ALTITUDE 4.- ACUCARES 5.- AMIDO 6.- ACUCAR 7.- BROTACAO 8.- CLIMA 9.- CLIMA TEMPERADO 10.- CROMATOGRAFIA 11.- VARIEDADE 12.- FRIO 13.- FLORACAO 14.- GEMA 15.- SECA 16.- SACAROSE 17.- TRABALHO	$C_i = \frac{5,5}{(9+17)-5,5} = \frac{5,5}{20,5} = 0,26$
2	EINHARDT, Patrícia Milech; CORREA, Elísia Rodrigues e RASEIRA, Maria do Carmo B.. Comparação entre métodos para testar a viabilidade de pólen de pessegueiro. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 5-7. ISSN 0100-2945.	1. PÊSSEGO 2. CULTURA IN VITRO 3. GERMINAÇÃO 4. PÓLEN	1.- POLEN 2.- PESSEGO 3.- ACUCAR 4.- ACUCAR CRISTAL 5.- AGAR 6.- AGUA 7.- ANALISE 8.- METODO ESTATISTICO 9.- PRODUCAO VEGETAL 10.- CORANTE 11.- COR 12.- VARIEDADE 13.- FRUTIFICACAO 14.- GERMINACAO 15.- LABORATORIO 16.- MEIO DE CULTURA 17.- MICROSCOPIO 18.- METODO 19.- TRABALHO	$C_i = \frac{3}{(4+19)-3} = \frac{3}{20} = 0,15$
3	MARTINS, Leila e SILVA, Walter Rodrigues da. Comportamento fisiológico de sementes de tangerina (*Citrus reticulata* Blanco) submetidas à desidratação. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 8-10. ISSN 0100-2945.	1. TANGERINA 2. SEMENTE 3. DESIDRATAÇÃO 4. UMIDADE	1.- DESIDRATACAO 2.- TANGERINA 3.- AGUA 4.- AR 5.- EQUIPAMENTO 6.- EMERGENCIA 7.- GERMINACAO 8.- HIPOCOTILO 9.- PESQUISA 10.- PLANTULA 11.- RAIZ 12.- SEMENTE 13.- TRATAMENTO 14.- TAXA 15.- UMIDADE 16.- VELOCIDADE	$C_i = \frac{4}{(4+16)-4} = \frac{4}{16} = 0,25$
4	SALOMAO, Luiz Carlos Chamhum; SIQUEIRA, Dalmo Lopes de e PEREIRA, Marcio Eduardo Canto. Desenvolvimento do fruto da lichieira (*Litchi chinensis* Sonn.) 'Bengal'. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 11-13. ISSN 0100-2945.	1. LICHIA 2. INFLORESCÊNCIA 3. FENOLOGIA 4. FRUTIFICAÇÃO 5. MATURAÇÃO 6. COLHEITA	1.- FRUTO 2.- COMPRIMENTO 3.- CRESCIMENTO 4.- COLHEITA 5.- COR 6.- DIAMETRO 7.- FRUTIFICACAO 8.- INFLORESCENCIA 9.- MATERIA SECA 10.- MATURACAO 11.- PERICARPO 12.- PELO 13.- SENESCENCIA 14.- SEMENTE 15.- SECA	$C_i = \frac{4}{(6+15)-4} = \frac{4}{17} = 0,23$

5	GARCIA-PEREZ, Eliseo e MARTINS, Antonio Baldo Geraldo. Florescimento e frutificação de lichieiras em função do anelamento de ramos. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 14-17. ISSN 0100-2945.	1. LICHIA 2. FRUTIFICAÇÃO 3. FRUTO 4. MATURAÇÃO	1.- ANELAGEM 2.- FLORACAO 3.- FRUTIFICACAO 4.- ARVORE 5.- COMPRIMENTO 6.- COLHEITA 7.- DIAMETRO	8.- EPOCA DE COLHEITA 9.- FRUTO 10.- IDADE 11.- MASSA 12.- PRODUCAO 13.- RENDIMENTO	$C_i = \frac{2}{(4+13)-2} = \frac{2}{15} = 0{,}13$
6	FRANZON, Rodrigo Cezar e RASEIRA, Maria do Carmo Bassols. Germinação *in vitro* e armazenamento do pólen de *Eugenia involucrata* DC (Myrtaceae). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 18-20. ISSN 0100-2945.	1. EUGENIA INVOLUCRATA 2. MYRTACEAE 3. CULTURA IN VITRO 4. GERMINAÇÃO 5. PÓLEN 6. ARMAZENAMENTO 7. SUL	1.- ARMAZENAMENTO 2.- EUGENIA 3.- EUGENIA INVOLUCRATA 4.- GERMINACAO 5.- POLEN 6.- ACUCAR 7.- AGAR 8.- AGUA	9.- BORO 10.- PRODUCAO VEGETAL 11.- INCUBACAO 12.- MYRTACEAE 13.- PERDA 14.- TRABALHO	$C_i = \frac{5}{(7+14)-5} = \frac{5}{16} = 0{,}31$
7	NEGREIROS, Jacson Rondinelli da Silva et al. Influência do estádio de maturação e do armazenamento pós-colheita na germinação e desenvolvimento inicial do maracujazeiro-amarelo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 21-24. ISSN 0100-2945.	1. MARACUJÁ 2. GERMINAÇÃO 3. MATURAÇÃO 4. PÓS-COLHEITA 5. ARMAZENAMENTO 6. FRUTO	1.- ARMAZENAMENTO 2.- GERMINACAO 3.- MATURACAO 4.- MARACUJA 5.- AREIA 6.- AREIA FINA 7.- COR 8.- EXTRACAO 9.- FRUTO	10.- PRODUCAO 11.- PELO 12.- POS-COLHEITA 13.- QUALIDADE 14.- TRABALHO 15.- TEMPERATURA 16.- UNIVERSIDADE 17.- UNIVERSIDADE FEDERAL	$C_i = \frac{6}{(6+17)-6} = \frac{6}{17} = 0{,}35$
8	BARROS, Daniella Inácio et al. Métodos de extração de sementes de mangaba visando à qualidade fisiológica. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 25-27. ISSN 0100-2945.	1. MANGABA 2. SEMENTE 3. TESTE DE VIGOR 4. EXTRAÇÃO 5. QUALIDADE	1.- EXTRACAO 2.- MANGABA 3.- QUALIDADE 4.- AREIA 5.- CONDUTIVIDADE ELETRICA 6.- EMERGENCIA 7.- GERMINACAO	8.- METODO 9.- MASSA 10.- SECA 11.- SEMENTE 12.- TRABALHO 13.- UMIDADE	$C_i = \frac{4}{(5+13)-4} = \frac{4}{14} = 0{,}28$

9	VERISSIMO, Valtair et al. Níveis de cálcio e boro de gemas florais de pereira (*Pyrus* sp.) no sul do Brasil. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 28-31. ISSN 0100-2945.	1. PERA 2. GEMA 3. NUTRIÇÃO VEGETAL 4. ANÁLISE FOLIAR 5. CÁLCIO 6. BORO 7. MACROELEMENTO 8. MICROELEMENTO 9. RIO GRANDE DO SUL	1.- BORO 2.- CALCIO 3.- ACLIMATACAO 4.- VARIEDADE 5.- CLIMA 6.- CLIMA TEMPERADO 7.- ABASTECIMENTO 8.- ESTACAO EXPERIMENTAL 9.- EMPRESA	10.- FLORACAO 11.- GEMA 12.- INVERNO 13.- METODO 14.- OUTONO 15.- PESQUISA 16.- PELO 17.- TRABALHO	$C_i = \frac{3}{(9+17)-3} = \frac{3}{23} = 0{,}13$
10	CITADIN, Idemir et al. Uso de cianamida hidrogenada e óleo mineral na floração, brotação e produção do pessegueiro 'Chiripá'. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 32-35. ISSN 0100-2945.	1. PÊSSEGO 2. DORMÊNCIA 3. BROTAÇÃO 4. FLORAÇÃO 5. PRODUTO QUÍMICO 6. TRATAMENTO 7. CLIMA 8. FRIO 9. ÓLEO MINERAL	1.- BROTACAO 2.- FLORACAO 3.- MINERAL 4.- OLEO 5.- OLEO MINERAL 6.- PRODUCAO 7.- PESSEGO 8.- COMBINADA 9.- COLHEITA	10.- VARIEDADE 11.- DORMENCIA 12.- FRIO 13.- FRUTIFICACAO 14.- INVERNO 15.- PATO 16.- PLANTA 17.- PLANTIO 18.- TRABALHO	$C_i = \frac{6}{(9+18)-6} = \frac{6}{21} = 0{,}28$
11	SANTOS, Adriana Ferreira dos; SILVA, Silvanda de Melo e ALVES, Ricardo Elesbão. Armazenamento de pitanga sob atmosfera modificada e refrigeração: I-transformações químicas em pós-colheita. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 36-41. ISSN 0100-2945.	1. PITANGA 2. VITAMINA C 3. CLOROFILA 4. CAROTENÓIDE 5. PÓS-COLHEITA 6. ARMAZENAMENTO 7. REFRIGERAÇÃO 8. PROPRIEDADE FÍSICO-QUÍMICA	1.- ARMAZENAMENTO 2.- ATMOSFERA 3.- PITANGA 4.- POS-COLHEITA 5.- REFRIGERACAO 6.- ACIDEZ 7.- ACUCARES	8.- CLOROFILA 9.- MATURACAO 10.- MANUTENCAO 11.- TEMPERATURA 12.- TAXA 13.- VITAMINA	$C_i = \frac{5{,}5}{(8+13)-5{,}5} = \frac{5{,}5}{15{,}5} = 0{,}35$
12	SANTOS, Adriana Ferreira dos; SILVA, Silvanda de Melo; MENDONCA, Rejane Maria Nunes e FILGUEIRAS, Heloísa Almeida Cunha. Armazenamento de pitangas sob atmosfera modificada e refrigeração: II - qualidade e conservação pós-colheita. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 42-45. ISSN 0100-2945.	1. PITANGA 2. PÓS-COLHEITA 3. ARMAZENAMENTO 4. PRESERVAÇÃO DE ALIMENTO 5. REFRIGERAÇÃO 6. MATURAÇÃO 7. TEMPERATURA	1.- ARMAZENAMENTO 2.- ATMOSFERA 3.- CONSERVACAO 4.- POS-COLHEITA 5.- QUALIDADE 6.- REFRIGERACAO	7.- MATURACAO 8.- MASSA 9.- PIGMENTACAO 10.- PODRIDAO 11.- TRABALHO	$C_i = \frac{4}{(7+11)-4} = \frac{4}{14} = 0{,}28$
13	AMARANTE, Cassandro Vidal Talamini do; CHAVES, Daniela Vieira e ERNANI, Paulo Roberto. Composição mineral e severidade de "bitter pit" em maçãs 'Catarina'. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 51-54. ISSN 0100-2945.	1. MAÇÃ 2. DISTÚRBIO FISIOLÓGICO 3. NUTRIÇÃO VEGETAL 4. PÓS-COLHEITA	1.- MINERAL 2.- ANALISE 3.- METODO ESTATISTICO 4.- CASCA 5.- VARIEDADE 6.- DISTURBIO 7.- DISTURBIO FISIOLOGICO	8.- FRUTO 9.- MATURACAO 10.- POMAR 11.- POLPA 12.- POS-COLHEITA	$C_i = \frac{2}{(4+12)-2} = \frac{2}{14} = 0{,}14$

14	LIMA, Maria Auxiliadora Coêlho de; SILVA, Adriane Luciana da; AZEVEDO, Suellen Soraia Nunes e SANTOS, Polyane de Sá. Tratamentos pós-colheita com 1-metilciclopropeno em manga 'Tommy Atkins': efeito de doses e número de aplicações. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 64-68. ISSN 0100-2945.	1. MANGA 2. PÓS-COLHEITA 3. ARMAZENAMENTO 4. MATURAÇÃO 5. POLPA DE FRUTA 6. ETILENO	1.- MANGA 2.- POS-COLHEITA 3.- ARMAZENAMENTO 4.- AMACIAMENTO 5.- MATURACAO 6.- CONSERVACAO 7.- COLHEITA	8.- DOSE 9.- ETILENO 10.- POLPA 11.- REFRIGERACAO 12.- TEMPERATURA	$C_i = \frac{5,5}{(6+12)-5,5} = \frac{5,5}{12,5} = 0,44$
15	RAGOZO, Carlos Renato Alves; LEONEL, Sarita e CROCCI, Adalberto José. Adubação verde em pomar cítrico. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 69-72. ISSN 0100-2945.	1. LARANJA 2. ADUBAÇÃO VERDE 3. CRESCIMENTO 4. PRODUÇÃO VEGETAL	1.- ADUBACAO 2.- ADUBACAO VERDE 3.- POMAR 4.- BIOMASSA 5.- CITRICULTURA 6.- PRODUCAO VEGETAL 7.- COLHEITA 8.- CRESCIMENTO 9.- ESTATISTICA 10.- LARANJA 11.- LIMAO 12.- LABE-LABE	13.- MATERIA SECA 14.- PESO 15.- PRODUTIVIDADE 16.- PRODUCAO 17.- QUALIDADE 18.- RENDIMENTO 19.- SEMEADURA 20.- SECA 21.- SUCO 22.- TRABALHO	$C_i = \frac{4}{(4+22)-4} = \frac{4}{22} = 0,18$
16	FAGUNDES, Angela Fuentes; DABUL, Audrei Nisio Gebieluca e AYUB, Ricardo Antonio. Aminoethoxivinilglicina no controle do amadurecimento de frutos de caqui cv. Fuyu. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 73-75. ISSN 0100-2945.	1. CAQUI 2. PÓS-COLHEITA 3. QUALIDADE 4. MATURAÇÃO TARDIA 5. ARMAZENAMENTO 6. REFRIGERAÇÃO	1.- MATURACAO 2.- CAQUI 3.- AGUA 4.- ADESIVO 5.- ACIDEZ 6.- DOSE 7.- ESPALHANTE 8.- ETILENO 9.- FRIO 10.- FISIOLOGIA	11.- FRUTO 12.- POS-COLHEITA 13.- PH 14.- POLPA 15.- QUALIDADE 16.- SOLUCAO 17.- TEMPERATURA	$C_i = \frac{3,5}{(6+17)-3,5} = \frac{3,5}{19,5} = 0,17$
17	FIGUEIREDO, José Orlando de et al. Comportamento de 16 porta-enxertos para o tangor Murcott na região de Itirapina-SP. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 76-78. ISSN 0100-2945.	1. LARANJA 2. LIMÃO 3. TANGOR MURCOTT 4. PORTA ENXERTO 5. QUALIDADE 6. SÃO PAULO	1.- TANGOR MURCOTT 2.- CLONE 3.- FAZENDA 4.- FRUTO 5.- LARANJA 6.- LIMAO	7.- CLONE NUCELAR 8.- QUALIDADE 9.- RAIO 10.- TANGERINA	$C_i = \frac{4}{(6+10)-4} = \frac{4}{12} = 0,33$

18	YAMANISHI, Osvaldo Kiyoshi et al. Comportamento do mamoeiro Sekati nas condições do oeste da Bahia. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 79-82. ISSN 0100-2945.	1. MAMÃO 2. CARACTERISTICAS AGRONÔMICAS 3. FENÓTIPO 4. QUALIDADE 5. PÓS-COLHEITA 6. PROPRIEDADE FÍSICO-QUÍMICA	1.- MAMAO 2.- ALTURA 3.- ACIDO 4.- CAULE 5.- CLOROFILA 6.- COMPRIMENTO 7.- DIAMETRO 8.- ESPESSURA 9.- HIBRIDO 10.- NERVURA 11.- PRIMAVERA	12.- PLANTA 13.- PESO 14.- POLPA 15.- PH 16.- POS-COLHEITA 17.- QUALIDADE 18.- REFRIGERACAO 19.- TEMPERATURA 20.- VARIEDADE 21.- VERAO	$C_i = \frac{3}{(6+21)-3} = \frac{3}{24} = 0{,}12$
19	LIMA, Rosiane de Lourdes Silva de; SIQUEIRA, Dalmo Lopes de; WEBER, Olmar Baller e CAZETTA, Jairo Osvaldo. Comprimento de estacas e parte do ramo na formação de mudas de aceroleira. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 83-86. ISSN 0100-2945.	1. ACEROLA 2. MUDA 3. REPRODUÇÃO VEGETAL 4. ENRAIZAMENTO DE ESTACA 5. BROTAÇÃO 6. TESTE DE VIGOR	1.- COMPRIMENTO 2.- RAMO 3.- INDUSTRIA AGRICOLA 4.- ARROZ 5.- BROTACAO 6.- ESTUFA 7.- CASCA 8.- CASCA DE ARROZ 9.- ESTACA 10.- IDADE	11.- MISTURA 12.- MATERIA SECA 13.- PULVERIZACAO 14.- PARCELA 15.- PARTE AEREA 16.- PRODUCAO 17.- SECA 18.- SISTEMA RADICULAR 19.- TAMANHO 20.- VEGETACAO	$C_i = \frac{1}{(6+20)-1} = \frac{1}{25} = 0{,}04$
20	SOUZA, Laercio Duarte et al. Distribuição das raízes dos citros em função da profundidade da cova de plantio em Latossolo Amarelo dos Tabuleiros Costeiros. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 87-91. ISSN 0100-2945.	1. FRUTA CÍTRICA 2. PRÁTICA CULTURAL 3. RAÍZ 4. DIÂMETRO 5. LATOSSOLO 6. ANÁLISE DO SOLO 7. FÍSICA DO SOLO 8. QUÍMICA DO SOLO	1.- FRUTA CITRICA 2.- COVA 3.- ABASTECIMENTO 4.- LATOSSOLO 5.- LATOSSOLO AMARELO 6.- PROFUNDIDADE 7.- PLANTIO 8.- AGUA 9.- CITRICULTURA	10.- DISPONIBILIDADE DE AGUA 11.- FRUTICULTURA 12.- LARANJA 13.- LIMAO 14.- SOLO 15.- VOLUME	$C_i = \frac{2{,}5}{(8+15)-2{,}5} = \frac{2{,}5}{20{,}5} = 0{,}12$
21	CAVICHIOLI, José Carlos et al. Florescimento e frutificação do maracujazeiro-amarelo submetido à iluminação artificial, irrigação e sombreamento. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 92-96. ISSN 0100-2945.	1. MARACUJÁ 2. ILUMINAÇÃO ARTIFICIAL 3. CLIMATOLOGIA 4. ENTRESSAFRA 5. FLORAÇÃO 6. IRRIGAÇÃO	1.- FLORACAO 2.- FRUTIFICACAO 3.- ILUMINACAO 4.- ILUMINACAO ARTIFICIAL 5.- IRRIGACAO 6.- SOMBREAMENTO 7.- AR 8.- ESCOLA 9.- ENTRESSAFRA	10.- MARACUJA 11.- PRODUCAO 12.- PRODUTIVIDADE 13.- SOLO 14.- TEMPERATURA 15.- TRABALHO 16.- TRATAMENTO	$C_i = \frac{5}{(6+16)-5} = \frac{5}{17} = 0{,}29$
22	JUNQUEIRA, Nilton Tadeu Vilela et al. Reação a doenças e produtividade de um clone de maracujazeiro-azedo propagado por estaquia e enxertia em estacas herbáceas de Passiflora	1. MARACUJÁ 2. CLONE 3. PORTA ENXERTO 4. TUTORAMENTO	1.- CLONE 2.- ENXERTO 3.- MARACUJA 4.- PRODUTIVIDADE	9.- IRRIGACAO 10.- LATOSSOLO 11.- PLANTIO 12.- RESISTENCIA	$C_i = \frac{3{,}5}{(10+14)-3{,}5} = \frac{3{,}5}{20{,}5} = 0{,}17$

	silvestre. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 97-100. ISSN 0100-2945.	5. REPRODUÇÃO ASSEXUADA 6. SOLO 7. FUNGO 8. VÍRUS 9. ANÁLISE FÍSICA 10. VARIEDADE RESISTENTE	5.- ALTURA 6.- CAMPO 7.- DIAMETRO 8.- ESPECIE	13.- SOLO 14.- TRABALHO	
23	SILVA, Edicléia Aparecida da; BOLIANI, Aparecida Conceição e CORREA, Luiz de Souza. Avaliação de cultivares de bananeira (*Musa* sp) na região de Selvíria-MS. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 101-103. ISSN 0100-2945.	1. BANANA 2. VARIEDADE 3. PRÁTICA CULTURAL 4. FRUTO 5. PRODUTIVIDADE 6. TAMANHO 7. MELHORAMENTO GENÉTICO VEGETAL 8. MATO GROSSO DO SUL	1.- BANANA 2.- VARIEDADE 3.- COLHEITA 4.- CACHO 5.- COMPRIMENTO 6.- DIAMETRO 7.- FLORACAO	8.- MATO 9.- MARMELO 10.- PRODUCAO 11.- PLANTIO 12.- TRABALHO	$C_i = \frac{2}{(8+12)-2} = \frac{2}{18} = 0{,}11$
24	REVERS, Luís Fernando et al. Uso prático de marcadores moleculares para seleção assistida no melhoramento de uvas de mesa apirênicas. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 104-108. ISSN 0100-2945.	1. UVA 2. MELHORAMENTO GENÉTICO VEGETAL 3. FENÓTIPO 4. MARCADOR MOLECULAR 5. MATO GROSSO DO SUL	1.- MELHORAMENTO 2.- SELECAO 3.- FENOTIPO 4.- GENETICA 5.- GENE 6.- MARCADOR MOLECULAR 7.- PROGENIE	8.- DISSEMINACAO SELETIVA DA INFORMACAO 9.- SEMENTE 10.- TRABALHO 11.- UVA 12.- VINHO	$C_i = \frac{3{,}5}{(5+12)-3{,}5} = \frac{3{,}5}{13{,}5} = 0{,}25$
25	DAMATTO JUNIOR, Erval Rafael; BOAS, Roberto Lyra Villas; LEONEL, Sarita e FERNANDES, Dirceu Maximino. Avaliação nutricional em folhas de bananeira 'Prata-anã' adubadas com composto orgânico. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 109-112. ISSN 0100-2945.	1. BANANA 2. ADUBO ORGÂNICO 3. NUTRIÇÃO VEGETAL 4. ANÁLISE FOLIAR 5. POTÁSSIO	1.- BANANA 2.- COMPOSTO ORGANICO 3.- ANALISE 4.- ADUBACAO 5.- COLHEITA 6.- PRODUCAO VEGETAL 7.- ESTADO NUTRICIONAL	8.- FLORACAO 9.- NUTRICAO 10.- PRODUCAO 11.- POTASSIO 12.- PARCELA 13.- TRABALHO	$C_i = \frac{3}{(5+13)-3} = \frac{3}{15} = 0{,}20$

26	ROSA, Raul Castro Carriello et al. Doses de nitrogênio e potássio em fertirrigação em maracujazeiro-amarelo consorciado com coqueiro-anão verde, na região Norte Fluminense. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 113-116. ISSN 0100-2945.	1. MARACUJÁ 2. COCO 3. CONSORCIAÇÃO DE CULTURA 4. ADUBAÇÃO 5. NITROGÊNIO 6. POTÁSSIO 7. FERTIRRIGAÇÃO 8. ANÁLISE FOLIAR	1.- FERTIRRIGACAO 2.- MARACUJA 3.- NITROGENIO 4.- POTASSIO 5.- AGUA 6.- AGUA DE IRRIGACAO 7.- ADUBACAO 8.- AMOSTRAGEM 9.- FOLHA	10.- IRRIGACAO 11.- INJECAO 12.- MICROASPERSAO 13.- PRODUTIVIDADE 14.- PESO 15.- RAMO 16.- SOLO 17.- VAZAO	$C_i = \frac{5}{(8+17)-5} = \frac{5}{20} = 0{,}25$
27	COSTA, Raquel Silva; OLIVEIRA, Inez Vilar de Morais; MORO, Fabíola Vitti e MARTINS, Antônio Baldo Geraldo. Aspectos morfológicos e influência do tamanho da semente na germinação do jambo-vermelho. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 117-120. ISSN 0100-2945.	1. JAMBO 2. MORFOLOGIA VEGETAL 3. SEMENTE 4. GERMINAÇÃO 5. TESTE DE VIGOR	1.- GERMINACAO 2.- SEMENTE 3.- TAMANHO 4.- COMPRIMENTO 5.- ESPECIE 6.- PERICARPO 7.- EMERGENCIA 8.- FAMILIA 9.- FERRO	10.- FOSFORO 11.- FRUTO 12.- IDENTIFICACAO 13.- LARGURA 14.- MYRTACEAE 15.- PLANTULA 16.- RAIZ 17.- SEMEADURA 18.- TRABALHO	$C_i = \frac{2}{(5+18)-2} = \frac{2}{21} = 0{,}09$
28	FERREIRA, Gisela et al. Curva de absorção de água em sementes de atemóia (*Annona cherimola* Mill. x *Annona squamosa* L.) cv. Gefner. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 121-124. ISSN 0100-2945.	1. ATEMÓIA 2. SEMENTE 3. ABSORÇÃO DE ÁGUA 4. GERMINAÇÃO	1.- ABSORCAO 2.- ABSORCAO DE AGUA 3.- AGUA 4.- ATEMOIA 5.- ANNONA SQUAMOSA 6.- EMBEBICAO 7.- FILTRO	8.- GERMINACAO 9.- IMERSAO 10.- PAPEL 11.- PARCELA 12.- TRABALHO	$C_i = \frac{3}{(4+12)-3} = \frac{3}{13} = 0{,}23$
29	LIMA, Adelise de Almeida; CALDAS, Ranulfo Corrêa e SANTOS, Vanderlei da Silva. Germinação e crescimento de espécies de maracujá. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 125-127. ISSN 0100-2945.	1. MARACUJA 2. PROPAGAÇÃO VEGETATIVA 3. ENXERTO 4. BAHIA	1.- GERMINACAO 2.- MARACUJA 3.- ALTURA 4.- ESTUFA 5.- EMERGENCIA 6.- ENXERTO 7.- ESPECIE 8.- FRUTICULTURA 9.- MANDIOCA	10.- PASSIFLORA EDULIS 11.- PARCELA 12.- PLANTA 13.- PROPAGACAO VEGETATIVA 14.- TRABALHO 15.- VELOCIDADE 16.- VEGETACAO	$C_i = \frac{3}{(4+16)-3} = \frac{3}{17} = 0{,}17$
30	ROBERTO, Sérgio Ruffo; KAGUEYAMA, Marcel Hiroaki e SANTOS, Cristiano Ezequiel dos. Indução da brotação da macieira 'Eva' em região de baixa incidência de frio. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 128-130. ISSN 0100-2945	1. MAÇÃ 2. DORMÊNCIA 3. QUEBRA DA DORMÊNCIA 4. BROTAÇÃO INDUZIDA 5. PRODUTO QUÍMICO	1.- BROTACAO 2.- FRIO 3.- MACA 4.- DORMENCIA 5.- FRUTIFICACAO 6.- MINERAL	7.- OLEO 8.- OLEO MINERAL 9.- RAMO 10.- TRABALHO 11.- TRATAMENTO	$C_i = \frac{2{,}5}{(5+11)-2{,}5} = \frac{2{,}5}{13{,}5} = 0{,}18$

31	SILVA, Katiane Santiago et al. Patogenicidade causada pelo fungo *Colletotrichum gloeosporioides* (Penz) em diferentes espécies frutíferas. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 131-133. ISSN 0100-2945.	1. MANGA 2. MAMÃO 3. GOIABA 4. MARACUJÁ 5. DOENÇA DE PLANTA 6. FUNGO	1.- COLLETOTRICHUM GLOEOSPORIOIDES 2.- FUNGO 3.- PATOGENICIDADE 4.- PELO 5.- ANTRACNOSE 6.- PRODUCAO VEGETAL 7.- CRESCIMENTO 8.- GOIABA 9.- HOSPEDEIRO	10.- INOCULACAO 11.- MANGA 12.- MAMAO 13.- MARACUJA 14.- MICELIO 15.- ROSA 16.- TRABALHO	$C_i = \frac{5}{(6+16)-5} = \frac{5}{17} = 0{,}29$
32	CARNAUBA, Juliana Paiva et al. *Phytophthora palmivora*, agente da podridão de raiz e frutos de mamoeiro no Estado de Alagoas. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 134-135. ISSN 0100-2945.	1. MAMÃO 2. DORMÊNCIA DA SEMENTE 3. PODRIDÃO 4. PERDA 5. FUNGO 6. ALAGOAS	1.- MAMAO 2.- PHYTOPHTHORA 3.- PHYTOPHTHORA PALMIVORA 4.- PODRIDAO 5.- VARIEDADE	6.- DOENCA 7.- PRODUCAO 8.- PATOGENICIDADE 9.- PELO	$C_i = \frac{2}{(6+9)-2} = \frac{2}{13} = 0{,}15$
33	VIEIRA, Cássia Regina Yuriko Ide et al. Fertilidade de gemas de videiras 'Niagara Rosada' de acordo com o sistema de condução. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 136-138. ISSN 0100-2945.	1. UVA 2. GEMA 3. FERTILIDADE 4. INFLORESCÊNCIA 5. MATO GROSSO DO SUL	1.- FERTILIDADE 2.- GEMA 3.- PELO	4.- UVA	$C_i = \frac{3}{(5+4)-3} = \frac{3}{6} = 0{,}50$
34	DONATO, Sérgio Luiz Rodrigues et al. Comportamento de variedades e híbridos de bananeira (*Musa* spp.), em dois ciclos de produção no sudoeste da Bahia. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 139-144. ISSN 0100-2945.	1. BANANA 2. CARACTERISTICAS AGRONÔMICAS 3. GENÓTIPO 4. REPRODUÇÃO VEGETAL 5. TAXA DE CRESCIMENTO 6. COMPORTAMENTO DE VARIEDADE 7. HÍBRIDO	1.- BANANA 2.- PRODUCAO 3.- ALTURA 4.- COLHEITA 5.- CACHO 6.- COMPRIMENTO 7.- DIAMETRO 8.- ESPACAMENTO 9.- FRUTICULTURA 10.- FLORACAO 11.- FRUTO 12.- HIBRIDO	13.- IDENTIFICACAO 14.- MANDIOCA 15.- PLANTA 16.- PLANTIO 17.- PESO 18.- PRODUTIVIDADE 19.- QUALIDADE 20.- SELECAO 21.- TRABALHO 22.- VARIEDADE	$C_i = \frac{2{,}5}{(7+22)-2{,}5} = \frac{2{,}5}{26{,}5} = 0{,}09$
35	SOUZA, Paulo Sergio de et al. Reação de variedades e clones de laranjas a *Xylella fastidiosa*. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 145-147. ISSN 0100-2945.	1. LARANJA 2. DOENÇA DE PLANTA 3. BACTÉRIA 4. CLONE	1.- XYLELLA FASTIDIOSA 2.- BACTERIA 3.- CLOROSE 4.- FRUTA CITRICA 5.- DOENCA 6.- ENCOSTIA 7.- INOCULACAO	8.- METODO 9.- PARCELA 10.- PINTO 11.- TRABALHO 12.- VARIEDADE	$C_i = \frac{1{,}5}{(4+12)-1{,}5} = \frac{1{,}5}{14{,}5} = 0{,}10$

36	PEREIRA, Bruno Fernando Faria e CARVALHO, Sérgio Alves de. Métodos de forçamento de borbulhas e aplicação de cianamida hidrogenada para produção de mudas de laranja 'Valência' sobre citrumelo 'Swingle' em viveiro telado. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 151-153. ISSN 0100-2945.	1. LARANJA 2. PROPAGAÇÃO VEGETATIVA 3. ENXERTO 4. REGULADOR DE CRESCIMENTO 5. MUDA	1.- LARANJA 2.- PRODUCAO 3.- VIVEIRO 4.- ALTURA 5.- BROTACAO 6.- CRESCIMENTO	7.- COMPRIMENTO 8.- DIAMETRO 9.- ENXERTO 10.- CAULE	$C_i = \frac{2,5}{(5+10)-2,5} = \frac{2,5}{12,5} = 0,20$
37	WEBER, Olmar Baller et al. Adubação nitrogenada e potássica em bananeira 'Pacovan' (*musa* AAB, subgrupo prata) na chapada do Apodi, Estado do Ceará. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.1, pp. 154-157. ISSN 0100-2945.	1. BANANA 2. ADUBAÇÃO 3. FERTILIZANTE NITROGENADO 4. FERTILIZANTE POTÁSSICO 5. IRRIGAÇÃO 6. PRODUTIVIDADE 7. FRUTO 8. QUALIDADE 9. CEARÁ	1.- ADUBACAO 2.- BANANA 3.- ACUCARES 4.- ACIDEZ 5.- CLORETO DE POTASSIO 6.- CACHO 7.- DOSE 8.- ESTERCO 9.- FOSFATO 10.- IRRIGACAO	11.- NITROGENIO 12.- POTASSIO 13.- PRODUCAO 14.- PLANTIO 15.- PRODUTIVIDADE 16.- QUALIDADE 17.- TRABALHO 18.- UREIA	$C_i = \frac{5}{(9+18)-5} = \frac{5}{22} = 0,22$
38	ATAIDE, Elma Machado et al. Efeito do paclobutrazol e de ácido giberélico na indução floral do maracujazeiro-amarelo em condições de entressafra. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 160-163. ISSN 0100-2945.	1. MARACUJÁ 2. REGULADOR DE CRESCIMENTO 3. ÁCIDO GIBERÉLICO 4. FLORAÇÃO	1.- ACIDO 2.- ACIDO GIBERELICO 3.- ENTRESSAFRA 4.- FLORACAO 5.- INDUCAO 6.- ADESIVO 7.- COMPRIMENTO	8.- ESPALHANTE 9.- ENTRENOS 10.- PLANTA 11.- PARCELA 12.- SOLO	$C_i = \frac{2}{(4+12)-2} = \frac{2}{14} = 0,14$
39	SANTOS, Janaína Pereira dos e WAMSER, Anderson Fernando. Efeito do ensacamento de frutos sobre danos causados por fatores bióticos e abióticos em pomar orgânico de macieira. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 168-171. ISSN 0100-2945.	1. MAÇÃ 2. CONTROLE BIOLÓGICO 3. EMBALAGEM 4. PLÁSTICO 5. PRODUÇÃO ORGÂNICA	1.- MACA 2.- POMAR 3.- COLHEITA 4.- VARIEDADE 5.- CHUVA 6.- CUSTO 7.- EMPRESA 8.- EMBALAGEM	9.- PELE 10.- GRANIZO 11.- PELO 12.- PRODUCAO 13.- PRODUCAO ORGANICA 14.- SAFRA 15.- TRABALHO 16.- TRATAMENTO	$C_i = \frac{3}{(5+16)-3} = \frac{3}{18} = 0,16$
40	DAVOGLIO JUNIOR, Antonio Carlos; BORDIN, Ivan e NEVES, Carmen Silvia Vieira Janeiro. Sistema radicular e desenvolvimento de plantas cítricas provenientes de viveiro telado e aberto. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 172-175. ISSN 0100-2945.	1. FRUTA CÍTRICA 2. MUDA 3. RAÍZ 4. PROPAGAÇÃO VEGETATIVA 5. SÃO PAULO	1.- SISTEMA RADICULAR 2.- VIVEIRO 3.- FRUTA CITRICA 4.- DIAMETRO 5.- PELO	6.- PLANTIO 7.- TRABALHO 8.- VOLUME	$C_i = \frac{1}{(5+8)-1} = \frac{1}{12} = 0,08$

41	CARVALHO FILHO, Celso Duarte; HONORIO, Sylvio Luis e GIL, José Moure. Qualidade pós-colheita de cerejas cv. Ambrunés utilizando coberturas comestíveis. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 180-184. ISSN 0100-2945.	1. CEREJA 2. ARMAZENAMENTO 3. PÓS-COLHEITA 4. CERA 5. CARNAÚBA 6. PRESERVAÇÃO DE ALIMENTO	1.- POS-COLHEITA 2.- QUALIDADE 3.- ACIDEZ 4.- ARMAZENAMENTO 5.- CERA 6.- CARNAUBA 7.- CONSERVACAO 8.- COMERCIALIZACAO 9.- CONTAMINACAO 10.- CONTAMINACAO FUNGICA 11.- DETERIORACAO 12.- EMULSAO 13.- IMERSAO	14.- MATURACAO 15.- PULVERIZACAO 16.- PERDA 17.- PERDA DE PESO 18.- PESO 19.- PEDUNCULO 20.- PODRIDAO 21.- TRABALHO 22.- TRATAMENTO 23.- UMIDADE 24.- UMIDADE RELATIVA 25.- ZEINA	$C_i = \frac{4}{(6+25)-4} = \frac{4}{27} = 0{,}14$
42	MALGARIM, Marcelo Barbosa; CANTILLANO, Rufino Fernando Flores e COUTINHO, Enilton Fick. Sistemas e condições de colheita e armazenamento na qualidade de morangos cv. Camarosa. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 185-189. ISSN 0100-2945.	1. MORANGO 2. ARMAZENAMENTO 3. REFRIGERAÇÃO 4. QUALIDADE 5. RADIAÇÃO LUMINOSA 6. TEMPERATURA	1.- ARMAZENAMENTO 2.- COLHEITA 3.- QUALIDADE 4.- ATMOSFERA 5.- ACIDEZ 6.- ACIDO 7.- VITAMINA 8.- COMERCIALIZACAO 9.- COR 10.- RADIACAO LUMINOSA 11.- MADEIRA	12.- MASSA 13.- PRODUTOR 14.- POLPA 15.- POLIETILENO 16.- PERDA 17.- REDUCAO 18.- SIMULACAO 19.- TRABALHO 20.- TEMPERATURA	$C_i = \frac{4}{(6+20)-4} = \frac{4}{22} = 0{,}18$
43	MOTA, Wagner Ferreira da et al. Uso de cera de carnaúba e saco plástico poliolefínico na conservação pós-colheita do maracujá-amarelo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 190-193. ISSN 0100-2945.	1. MARACUJA 2. ARMAZENAMENTO 3. ATMOSFERA 4. EMBALAGEM 5. PLÁSTICO 6. PÓS-COLHEITA 7. PERDA 8. POLPA	1.- CERA 2.- CARNAUBA 3.- CONSERVACAO 4.- PLASTICO 5.- POS-COLHEITA 6.- AGUA 7.- ACIDEZ 8.- ARMAZENAMENTO 9.- ATMOSFERA 10.- COMERCIALIZACAO 11.- CASCA	12.- EMBALAGEM 13.- FRUTO 14.- IMERSAO 15.- MASSA 16.- PELO 17.- PERDA 18.- POLPA 19.- PERICARPO 20.- REDUCAO 21.- TRABALHO	$C_i = \frac{7}{(8+21)-7} = \frac{7}{22} = 0{,}31$

44	TOLEDO, Francisco Ricardo de; BARBOSA, José Carlos e YAMAMOTO, Pedro Takao. Distribuição espacial de *Toxoptera citricida* (Kirkaldy) (Hemiptera: Aphididae) na cultura de citros. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 194-198. ISSN 0100-2945.	1. LARANJA 2. PRAGA DE PLANTA 3. CONTROLE INTEGRADO 4. DISTRIBUIÇÃO GEOGRÁFICA	1.- PRODUCAO VEGETAL 2.- FRUTA CITRICA 3.- ABASTECIMENTO 4.- AMOSTRAGEM 5.- ALTURA 6.- COR 7.- DISPERSAO 8.- IDADE	9.- MANEJO 10.- CONTROLE INTEGRADO 11.- PLANTA 12.- PULGAO 13.- POPULACAO 14.- PRAGA 15.- VARIEDADE	$C_i = \frac{1,5}{(4+15)-1,5} = \frac{1,5}{17,5} = 0,08$
45	MONTESINO, Luiz Henrique; COELHO, Juliana Helena Carvalho; FELIPPE, Marcos Rogério e YAMAMOTO, Pedro Takao. Ingestão de seiva do xilema de laranjeiras 'Pêra' e 'Valência' (*Citrus sinensis* (L.) Osbeck) sadias e infectadas por *Xylella fastidiosa*, pelas cigarrinhas vetoras *Oncometopia facialis* e *Dilobopterus costalimai* (Hemiptera: Cicadellidae). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 199-204. ISSN 0100-2945.	1. LARANJA 2. XYLELLA FASTIDIOSA 3. BACTÉRIA 4. CIGARRINHA 5. PRAGA DE PLANTA	1.- INGESTAO 2.- XILEMA 3.- XYLELLA FASTIDIOSA 4.- ALIMENTACAO 5.- BACTERIA 6.- CLOROSE 7.- FRUTA CITRICA 8.- CAMPO 9.- CONSUMO 10.- CIGARRINHA 11.- HABITO ALIMENTAR	12.- EXCRECAO 13.- INOCULACAO 14.- LINHAGEM 15.- LIQUIDO 16.- PLANTA 17.- RAMO 18.- TAXA 19.- VOLUME 20.- VARIEDADE	$C_i = \frac{3}{(5+20)-3} = \frac{3}{22} = 0,13$
46	SILVA, Marcos Zatti da e OLIVEIRA, Carlos Amadeu Leite de. Seletividade de alguns agrotóxicos em uso na citricultura ao ácaro predador *Neoseiulus californicus* (McGregor) (Acari: Phytoseiidae). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 205-208. ISSN 0100-2945.	1. FRUTA CÍTRICA 2. PRAGA DE PLANTA 3. CONTROLE INTEGRADO 4. CONTROLE BIOLÓGICO 5. ÁCARO	1.- ACARO 2.- CITRICULTURA 3.- PREDADOR 4.- FRUTA CITRICA 5.- PRODUCAO VEGETAL 6.- CONTROLE BIOLOGICO 7.- ENXOFRE 8.- LABORATORIO	9.- METODO 10.- MORTALIDADE 11.- MANEJO 12.- CONTROLE INTEGRADO 13.- OXIDO 14.- PERA 15.- VARIEDADE	$C_i = \frac{4}{(5+15)-4} = \frac{4}{16} = 0,25$
47	MAIA, Ozana Maria de Andrade e OLIVEIRA, Carlos Amadeu Leite de. Suscetibilidade de cercas-vivas, quebra-ventos e plantas invasoras ao vírus da leprose e sua transmissão para laranjeiras por *Brevipalpus phoenicis* (Geijskes) (Acari: Tenuipalpidae). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 209-213. ISSN 0100-2945.	1. LARANJA 2. PLANTA HOSPEDEIRA 3. CERCA 4. ÁCARO 5. LEPROSE CÍTRICA 6. VÍRUS 7. CONTROLE BIOLÓGICO	1.- BREVIPALPUS PHOENICIS 2.- LEPROSE 3.- VIRUS 4.- ACARO 5.- ALGODAO 6.- BIXA ORELLANA 7.- BIDENS PILOSA 8.- COMMELINA 9.- ESTUFA	10.- GREVILLEA ROBUSTA 11.- MIMOSA 12.- MIMOSA CAESALPINIAEFOLIA 13.- PLANTA 14.- PLANTA HOSPEDEIRA 15.- SIDA 16.- VEGETACAO	$C_i = \frac{3,5}{(7+16)-3,5} = \frac{3,5}{19,5} = 0,17$
48	PEROSA, José Matheus; VIEIRA, Emerson Morais e NITZSCHE, Thomas. Cadeia produtiva da nêspera na região do alto Tietê: indicadores econômicos da produção e mercado atacadista. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 214-217. ISSN 0100-2945.	1. NESPERA 2. FRUTA TROPICAL 3. CADEIA PRODUTIVA 4. RENTABILIDADE 5. QUALIDADE 6. PREÇO 7. SÃO PAULO	1.- CADEIA PRODUTIVA 2.- MERCADO 3.- MERCADO ATACADISTA 4.- NESPERA 5.- PRODUCAO 6.- CRESCIMENTO	7.- CAMPO 8.- DIA DE CAMPO 9.- PESQUISA 10.- RENTABILIDADE 11.- TRABALHO	$C_i = \frac{3}{(7+11)-3} = \frac{3}{15} = 0,20$

49	ALMEIDA, Gabriel Vicente Bitencourt de e DURIGAN, José Fernando. Relação entre as características químicas e o valor dos pêssegos comercializados pelo sistema veiling frutas Holambra em Paranapanema-SP. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 218-221. ISSN 0100-2945.	1. PÊSSEGO 2. PROPRIEDADE ORGANOLÉPTICA 3. COMPOSIÇÃO QUÍMICA 4. COMERCIALIZAÇÃO 5. PREÇO 6. SÃO PAULO	1.- PELO 2.- ACIDEZ 3.- COMERCIALIZACAO 4.- CONSUMIDOR 5.- CONSUMO 6.- COOPERATIVA 7.- VARIEDADE 8.- DEMANDA 9.- ENTREPOSTO 10.- FRUTO	11.- LEILAO 12.- MELHORAMENTO 13.- PESSEGO 14.- PRODUTOR 15.- QUALIDADE 16.- RECEITA 17.- TRABALHO	$C_i = \frac{2}{(6+17)-2} = \frac{2}{21} = 0{,}09$
50	PIO, Rose Mary et al. Características da variedade Fremont quando comparadas com as das tangerinas 'Ponkan' e 'Clementina Nules'. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 222-226. ISSN 0100-2945.	1. TANGERINA 2. PONKAN 3. CARACTERÍSTICAS AGRONÔMICAS 4. HÍBRIDO 5. QUALIDADE 6. MERCADO 7. PROPRIEDADE FÍSICO-QUÍMICA	1.- VARIEDADE 2.- FRUTA CITRICA 3.- CONSUMO 4.- COLHEITA 5.- ENTRESSAFRA 6.- FRUTA	7.- MATURACAO 8.- QUALIDADE 9.- SAFRA 10.- TANGERINA	$C_i = \frac{2}{(7+10)-2} = \frac{2}{15} = 0{,}13$
51	MAYER, Newton Alex; PEREIRA, Fernando Mendes e KOBA, Valter Yoshio. Desenvolvimento inicial no campo de pessegueiros 'Aurora-1' enxertados em clones de umezeiro e 'Okinawa' propagados por estacas herbáceas. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 231-235. ISSN 0100-2945.	1. PÊSSEGO 2. PROPAGAÇÃO VEGETATIVA 3. PORTA ENXERTO 4. CLONE	1.- CAMPO 2.- BROTACAO 3.- CLONE 4.- ESPACAMENTO 5.- ENXERTO 6.- FLORACAO	7.- IRRIGACAO 8.- PARCELA 9.- PLANTA 10.- PESSEGO	$C_i = \frac{2{,}5}{(4+10)-2{,}5} = \frac{2{,}5}{11{,}5} = 0{,}21$
52	KREUZ, Carlos Leomar; SOUZA, Alceu e PETRI, José Luiz. Impacto da intensificação da densidade de plantio na rentabilidade em duas cultivares de macieira em Fraiburgo-SC. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 240-243. ISSN 0100-2945.	1. MAÇÃ 2. CUSTO DE PRODUÇÃO 3. RENTABILIDADE 4. AGRONEGÓCIO 5. PREÇO 6. SANTA CATARINA	1.- DENSIDADE DE PLANTIO 2.- MACA 3.- PLANTIO 4.- RENTABILIDADE 5.- AGRONEGOCIO	6.- ANALISE 7.- VARIEDADE 8.- PRODUCAO	$C_i = \frac{3}{(6+8)-3} = \frac{3}{11} = 0{,}27$
53	FERRARO, Aline Enila; PIO, Rose Mary e AZEVEDO, Fernando Alves de. Influência da polinização com variedades de laranja-doce sobre o número de sementes de tangelo Nova. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 244-246. ISSN 0100-2945.	1. TANGERINA 2. PROPRIEDADE ORGANOLÉPTICA 3. REPRODUÇÃO VEGETAL 4. POLINIZAÇÃO 5. SÃO PAULO	1.- POLINIZACAO 2.- BONITO 3.- FRUTA CITRICA 4.- ESTATISTICA 5.- FRUTO 6.- ISOLAMENTO 7.- LARANJA	8.- LARANJA PERA 9.- PERA 10.- TRABALHO 11.- TRATAMENTO 12.- TANGERINA	$C_i = \frac{2}{(5+12)-2} = \frac{2}{15} = 0{,}13$
54	RIBEIRO, Rafael Vasconcelos; MACHADO, Eduardo Caruso e BRUNINI, Orivaldo. Ocorrência de condições ambientais para a indução do florescimento de laranjeiras no Estado de São Paulo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 247-253. ISSN 0100-2945	1. LARANJA 2. CLIMA 3. TEMPERATURA 4. UMIDADE 5. FLORAÇÃO 6. INDUÇÃO	1.- FLORACAO 2.- INDUCAO 3.- CONDICAO AMBIENTAL 4.- BALANCO HIDRICO 5.- FRUTA CITRICA 6.- DEFICIENCIA	8.- FRIO 9.- JAU 10.- PELO 11.- RIO 12.- SECA 13.- TEMPERATURA	$C_i = \frac{3}{(6+13)-3} = \frac{3}{16} = 0{,}18$

			7.- DEFICIENCIA HIDRICA		
55	MANZONI, Cristiane Gindri et al. Seletividade de agrotóxicos recomendados na produção integrada da maçã a *Trichogramma pretiosum* Riley, 1879 (Hym.: Trichogrammatidae) em condições de laboratório. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 254-257. ISSN 0100-2945.	1. MAÇÃ 2. PRAGA DE PLANTA 3. INSETO 4. AGROTÓXICO 5. CONTROLE INTEGRADO 6. TRICHOGRAMMA SP 7. INSETO PARA CONTROLE BIOLÓGICO 8. CONTROLE QUÍMICO	1.- LABORATORIO 2.- MACA 3.- PRODUCAO 4.- ESTUFA 5.- CAMPO 6.- DIA DE CAMPO	7.- FORMULACAO 8.- PRODUTO 9.- TOXIDEZ 10.- VEGETACAO	$C_i = \frac{1}{(8+10)-1} = \frac{1}{17} = 0{,}05$
56	REGINA, Murilo de Albuquerque et al. Avaliação de híbridos de videira destinados à elaboração de vinhos brancos em Caldas, Minas Gerais. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 262-266. ISSN 0100-2945.	1. UVA BRANCA 2. COMPORTAMENTO DE VARIEDADE 3. HÍBRIDO 4. BROTAÇÃO	1.- UVA 2.- ANTRACNOSE 3.- ACIDEZ 4.- ACLIMATACAO 5.- BROTACAO 6.- COLHEITA	7.- FLORACAO 8.- MATURACAO 9.- MILDIO 10.- PRODUCAO 11.- QUALIDADE	$C_i = \frac{1{,}5}{(4+11)-1{,}5} = \frac{1{,}5}{13{,}5} = 0{,}11$
57	FARIA, Gláucia Amorim et al. Efeito da sacarose e sorbitol na conservação *in vitro* de *Passiflora giberti* N. E. Brown. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 267-270. ISSN 0100-2945.	1. MARACUJÁ 2. SACAROSE 3. CULTURA IN VITRO 4. PASSIFLORACEAE 5. CRESCIMENTO 6. CARBONO	1.- CONSERVACAO 2.- SACAROSE 3.- COMPRIMENTO 4.- COR 5.- CRESCIMENTO 6.- PRODUCAO VEGETAL 7.- CARBONO	8.- INCUBACAO 9.- MARACUJA 10.- MEIO DE CULTURA 11.- PELO 12.- TRABALHO 13.- TEMPERATURA	$C_i = \frac{4}{(6+13)-4} = \frac{4}{15} = 0{,}26$
58	SOUZA, Bianca Sarzi de; DURIGAN, José Fernando; DONADON, Juliana Rodrigues e SOUZA, Paulo Sergio de. Mangas minimamente processadas amadurecidas naturalmente ou com etileno e armazenadas em diferentes embalagens. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 271-275. ISSN 0100-2945.	1. MANGA 2. PROCESSAMENTO 3. ARMAZENAMENTO 4. EMBALAGEM 5. ETILENO 6. POLIETILENO	1.- ETILENO 2.- ACIDO 3.- VITAMINA 4.- ACIDEZ 5.- AMIDO 6.- ARMAZENAMENTO 7.- COR 8.- COMERCIALIZACAO 9.- DETERGENTE 10.- EMBALAGEM	11.- MANGA 12.- MATURACAO 13.- POLPA 14.- PH 15.- PEROXIDASE 16.- QUALIDADE 17.- RESISTENCIA 18.- REDUCAO	$C_i = \frac{4}{(6+18)-4} = \frac{4}{20} = 0{,}20$

59	SOUZA, Paulo Vítor Dutra de; CARNIEL, Edgar e FOCHESATO, Mário Luís. Efeito da composição do substrato no enraizamento de estacas de maracujazeiro azedo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 276-279. ISSN 0100-2945.	1. MARACUJA 2. PROPAGAÇÃO VEGETATIVA 3. ESTACA 4. CULTURA DE TECIDO 5. PROPRIEDADE FÍSICO-QUÍMICA 6. ENRAIZAMENTO 7. SUBSTRATO DE CULTURA	1.- ENRAIZAMENTO 2.- MARACUJA 3.- ARROZ 4.- AGUA 5.- CASCA 6.- CASCA DE ARROZ 7.- ESTUFA 8.- COMPRIMENTO 9.- ESTACA 10.- MICROASPERSAO	11.- PARCELA 12.- PH 13.- PELO 14.- SOLUCAO 15.- SECA 16.- VERMICULITA 17.- VEGETACAO 18.- VOLUME	$C_i = \frac{3}{(7+18)-3} = \frac{3}{22} = 0{,}13$
60	COSTA, Frederico Henrique da Silva; PEREIRA, Jonny Everson Scherwinski; PEREIRA, Maria Aparecida Alves e OLIVEIRA, Janiffe Peres de. Efeito da interação entre carvão ativado e N6-benzilaminopurina na propagação *in vitro* de bananeira, cv. Grand Naine (AAA). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 280-283. ISSN 0100-2945.	1. BANANA 2. CULTURA IN VITRO 3. CARVÃO 4. PROPAGAÇÃO VEGETATIVA 5. COMPOSTO FENÓLICO 6. OXIDAÇÃO	1.- BANANA 2.- CARVAO 3.- AGAR 4.- PRODUCAO VEGETAL 5.- CRESCIMENTO 6.- COMPRIMENTO 7.- MEIO DE CULTURA	8.- MUSA SP 9.- MICROPROPAGACAO 10.- OXIDACAO 11.- TRABALHO 12.- TAXA	$C_i = \frac{3}{(6+12)-3} = \frac{3}{15} = 0{,}20$
61	BRAGA, Marcelo Fideles et al. Enraizamento de estacas de três espécies silvestres de *Passiflora*. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 284-288. ISSN 0100-2945.	1. MARACUJA 2. ENRAIZAMENTO DE ESTACA 3. PROPAGAÇÃO VEGETATIVA	1.- ENRAIZAMENTO 2.- CELULA 3.- MORTALIDADE 4.- MARACUJA	5.- PULVERIZACAO 6.- PLASTICO	$C_i = \frac{1{,}5}{(3+6)-1{,}5} = \frac{1{,}5}{7{,}5} = 0{,}20$
62	SILVA, Breno Marques da Silva e et al. Germinação de sementes e emergência de plântulas de *Oenocarpus minor* Mart. (Arecaceae). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 289-292. ISSN 0100-2945.	1. BACABA 2. SEMENTE 3. PALMEIRA OLEAGINOSA 4. GERMINAÇÃO 5. EMERGÊNCIA	1.- EMERGENCIA 2.- GERMINACAO 3.- AGUA 4.- AREIA 5.- ESTUFA 6.- PROFUNDIDADE 7.- PROFUNDIDADE DE SEMEADURA	8.- SEMEADURA 9.- TRABALHO 10.- TEMPERATURA 11.- VERMICULITA 12.- VEGETACAO	$C_i = \frac{2}{(5+12)-2} = \frac{2}{15} = 0{,}13$
63	SANTOS, Breno Régis et al. Micropropagação de pequizeiro (*Caryocar brasiliense* Camb.). *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 293-296. ISSN 0100-2945.	1. PEQUI 2. PROPAGAÇÃO VEGETATIVA 3. BROTAÇÃO INDUZIDA 4. ACLIMATAÇÃO 5. REGULADOR DE CRESCIMENTO	1.- MICROPROPAGACAO 2.- ACIDO 3.- ACIDO INDOLBUTIRICO 4.- BROTACAO 5.- CARVAO 6.- PRODUCAO VEGETAL 7.- DORMENCIA 8.- EXPLANTE	9.- INDUCAO 10.- POLIETILENO 11.- SOBREVIVENCIA 12.- TRABALHO 13.- TRATAMENTO 14.- TAXA	$C_i = \frac{0{,}5}{(5+14)-0{,}5} = \frac{0{,}5}{18{,}5} = 0{,}02$

64	TELLES, Charles Allan; BIASI, Luiz Antonio; MINDELLO NETO, Ubirajara Ribeiro e PETERS, Eduardo. Sobrevivência e crescimento de mudas de pessegueiro interenxertadas. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 297-300. ISSN 0100-2945.	1. PÊSSEGO 2. ENXERTO 3. PROPAGAÇÃO VEGETATIVA 4. CRESCIMENTO 5. MUDA	1.- CRESCIMENTO 2.- PESSEGO 3.- SOBREVIVENCIA 4.- COMPRIMENTO 5.- DIAMETRO 6.- ENXERTO 7.- PRODUCAO	8.- PARCELA 9.- POMAR 10.- TECNOLOGIA 11.- TRABALHO 12.- VIVEIRO	$C_i = \frac{3}{(5+12)-3} = \frac{3}{14} = 0{,}21$
65	BORGES, Ana Lúcia; CALDAS, Ranulfo Corrêa e LIMA, Adelise de Almeida. Doses e fontes de nitrogênio em fertirrigação no cultivo do maracujá-amarelo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 301-304. ISSN 0100-2945.	1. MARACUJÁ 2. NITROGÊNIO 3. FERTIRRIGAÇÃO 4. ABSORÇÃO DE NUTRIENTES	1.- FERTIRRIGACAO 2.- NITROGENIO 3.- ADUBACAO 4.- CLIMA 5.- CALCIO 6.- ESPACAMENTO 7.- FRUTO 8.- IRRIGACAO 9.- LATOSSOLO 10.- LATOSSOLO AMARELO 11.- MARACUJA 12.- MINERAL	13.- NUTRICAO 14.- NUTRIENTE 15.- NITRATO 16.- PRODUTIVIDADE 17.- PELO 18.- PRODUCAO 19.- QUALIDADE 20.- SOLO 21.- SUCO 22.- TRABALHO 23.- UREIA	$C_i = \frac{3{,}5}{(4+23)-3{,}5} = \frac{3{,}5}{23{,}5} = 0{,}14$
66	PRADO, Renato de Melo; NATALE, William e ROZANE, Danilo Eduardo. Níveis críticos de boro no solo e na planta para cultivo de mudas de maracujazeiro-amarelo. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 305-309. ISSN 0100-2945.	1. MARACUJÁ 2. MUDA 3. ADUBAÇÃO 4. BORO 5. ABSORÇÃO DE NUTRIENTES	1.- BORO 2.- PLANTA 3.- SOLO 4.- ACIDO 5.- ACIDO BORICO 6.- ALTURA 7.- AGUA 8.- ADUBACAO 9.- CAULE 10.- DIAMETRO 11.- DOSE 12.- LATOSSOLO	13.- LATOSSOLO VERMELHO 14.- MATERIA SECA 15.- MARACUJA 16.- MICROELEMENTO 17.- NUTRICAO 18.- PRODUCAO 19.- PLANTIO 20.- PARTE AEREA 21.- SECA 22.- VEGETACAO	$C_i = \frac{3}{(5+22)-3} = \frac{3}{24} = 0{,}12$
67	SILVA, Ricardo Alencar da et al. Qualidade de frutos do coqueiro-anão verde fertirrigado com nitrogênio e potássio. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 310-313. ISSN 0100-2945.	1. COCO 2. QUALIDADE 3. ADUBAÇÃO 4. NITROGÊNIO 5. POTÁSSIO 6. IRRIGAÇÃO 7. ÁGUA DE COCO	1.- NITROGENIO 2.- POTASSIO 3.- QUALIDADE 4.- AGUA 5.- AGUA DE COCO 6.- CONDUTIVIDADE ELETRICA 7.- COCO 8.- DOSE 9.- FERTIRRIGACAO	10.- IDADE 11.- MATRIZ 12.- POMAR 13.- PESO 14.- PH 15.- RIO 16.- VOLUME	$C_i = \frac{5}{(7+16)-5} = \frac{5}{18} = 0{,}27$
68	BORGES, Ana Lúcia; SILVA, Sebastião de Oliveira e; CALDAS, Ranulfo Corrêa e LEDO, Carlos Alberto da Silva. Teores foliares de nutrientes em genótipos de bananeira. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 314-318. ISSN 0100-2945.	1. BANANA 2. GENÓTIPO 3. ANÁLISE FOLIAR 4. NUTRIÇÃO VEGETAL 5. VALOR NUTRITIVO	1.- BANANA 2.- ANALISE FOLIAR 3.- ESTADO NUTRICIONAL 4.- FOLHA 5.- FRUTICULTURA	6.- GERMOPLASMA 7.- MANDIOCA 8.- PRODUCAO 9.- TRABALHO	$C_i = \frac{2}{(5+9)-2} = \frac{2}{12} = 0{,}16$

69	ALMEIDA, Gustavo Costa; VILAS BOAS, Eduardo Valério de Barros; RODRIGUES, Luiz José e PAULA, Nélio Ranieli Ferreira de. Atraso do amadurecimento de banana 'Maçã' pelo 1-MCP, aplicado previamente à refrigeração. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 319-321. ISSN 0100-2945.	1. BANANA MAÇÃ 2. ARMAZENAMENTO 3. MATURAÇÃO 4. TEMPERATURA 5. COMPOSIÇÃO QUÍMICA 6. QUALIDADE 7. PH 8. MINAS GERAIS	1.- MATURACAO 2.- BANANA 3.- PELO 4.- REFRIGERACAO 5.- ARMAZENAMENTO 6.- CASCA 7.- COMPOSICAO QUIMICA 8.- COMERCIALIZACAO 9.- EXPOSICAO	10.- POS-COLHEITA 11.- QUALIDADE 12.- QUIMICA 13.- TRABALHO 14.- TEMPERATURA 15.- UMIDADE 16.- UMIDADE RELATIVA	$C_i = \frac{5,5}{(8+16)-5,5} = \frac{5,5}{18,5} = 0,29$
70	FONSECA, Marcos José de Oliveira et al. Emissão de etileno e de CO2 em mamão 'Sunrise Solo' e 'Golden'. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 322-324. ISSN 0100-2945.	1. MAMÃO 2. PÓS-COLHEITA 3. ARMAZENAMENTO 4. RESPIRAÇÃO 5. MATURAÇÃO 6. TEMPERATURA	1.- ETILENO 2.- MAMAO 3.- ARMAZENAMENTO 4.- MATURACAO 5.- COR 6.- CASCA	7.- METABOLISMO 8.- POS-COLHEITA 9.- REFRIGERACAO 10.- RESPIRACAO 11.- TEMPERATURA	$C_i = \frac{6}{(6+11)-6} = \frac{6}{11} = 0,54$
71	BOLOGNA, Isabela Rodrigues e VITTI, Godofredo Cesar. Produção e qualidade de frutos de laranjeira 'Pêra' em função de fontes e doses de boro. *Rev. Bras. Frutic.* [online]. 2006, vol.28, n.2, pp. 328-330. ISSN 0100-2945	1. LARANJA 2. ADUBAÇÃO 3. BORO 4. ÁCIDO BORICO 5. MICROELEMENTO 6. NUTRIÇÃO VEGETAL 7. QUALIDADE	1.- BORO 2.- LARANJA 3.- PRODUCAO 4.- QUALIDADE 5.- ACIDO 6.- ACIDO BORICO 7.- CASCA 8.- DOSE 9.- DIAMETRO	10.- ESPESSURA 11.- FRUTO 12.- REDUCAO 13.- RENDIMENTO 14.- SOLO 15.- SUCO 16.- TRABALHO	$C_i = \frac{4}{(7+16)-4} = \frac{4}{23} = 0,17$
72	PINHEIRO, Ana Carla Marques; VILAS BOAS, Eduardo Valério de Barros; ALVES, Alessandra de Paiva e LA SELVA, Marcelo. Amadurecimento de bananas 'maçã' submetidas ao 1-metilciclopropeno (1-MCP). *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 1-4. ISSN 0100-2945.	1. BANANA 2. PÓS-COLHEITA 3. CASCA 4. TAXA 5. RESPIRAÇÃO 6. COR 7. QUALIDADE 8. PROPRIEDADE FÍSICO-QUÍMICA 9. MÉTODO ESTATÍSTICO	1.- MATURACAO 2.- ANALISE 3.- METODO ESTATISTICO 4.- ACUCARES 5.- BANANA 6.- COR	7.- CASCA 8.- PH 9.- POS-COLHEITA 10.- QUALIDADE 11.- TAXA	$C_i = \frac{7}{(9+11)-7} = \frac{7}{13} = 0,53$
73	SARTORI, Ivar Antonio et al. Efeito da poda, raleio de frutos e uso de fitorreguladores na produção de tangerinas (*Citrus deliciosa* Tenore) cv. montenegrina. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 5-10. ISSN 0100-2945.	1. TANGERINA 2. PODA 3. REGULADOR DE CRESCIMENTO 4. PRÁTICA CULTURAL 5. CONTROLE DE QUALIDADE	1.- PODA 2.- PRODUCAO 3.- RALEIO 4.- ACIDO 5.- ACIDO GIBERELICO 6.- FRUTA CITRICA 7.- CRESCIMENTO 8.- EMPRESA 9.- FRUTIFICACAO 10.- FLORACAO	11.- IDADE 12.- LARANJA 13.- MASSA 14.- POMAR 15.- PULVERIZACAO 16.- PARCELA 17.- PLANTA 18.- QUALIDADE 19.- TAMANHO	$C_i = \frac{2}{(5+19)-2} = \frac{2}{22} = 0,09$

74	MESQUITA, Marcos Antônio Machado et al. Caracterização de ambientes com alta ocorrência natural de araticum (*Annona crassiflora* Mart.) no Estado de Goiás. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 15-19. ISSN 0100-2945.	1. ARATICUM 2. MEIO AMBIENTE 3. ECOLOGIA VEGETAL 4. POPULAÇÃO DE PLANTA 5. GEOMORFOLOGIA 6. SOLO 7. DISTRIBUIÇÃO GEOGRÁFICA 8. LATOSSOLO 9. GOIÁS	1.- ARATICUM 2.- AREA BASAL 3.- CERRADO	4.- ESPECIE 5.- SOLO	$C_i = \frac{2}{(9+5)-2} = \frac{2}{12} = 0{,}16$
75	HOJO, Ronaldo Hissayuki et al. Caracterização fenológica da goiabeira 'Pedro Sato' sob diferentes épocas de poda. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 20-24. ISSN 0100-2945.	1. GOIABA 2. PODA 3. BROTAÇÃO 4. FLORAÇÃO 5. ETAPA DE DESENVOLVIMENTO DA PLANTA 6. CLIMA 7. TEMPERATURA 8. FENOLOGIA 9. MINAS GERAIS	1.- GOIABA 2.- PODA 3.- BROTACAO 4.- PRODUCAO VEGETAL 5.- COLHEITA 6.- FLORACAO 7.- FLOR 8.- FRUTO	9.- FENOLOGIA 10.- IDADE 11.- MATURACAO 12.- PRODUCAO 13.- PRODUTOR 14.- PLANTA	$C_i = \frac{5}{(9+14)-5} = \frac{5}{18} = 0{,}27$
76	OLIVEIRA, Lenaldo Muniz de et al. Efeito de citocininas na senescência e abscisão foliar durante o cultivo *in vitro* de *Annona glabra* L.. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 25-30. ISSN 0100-2945.	1. ARATICUM DO BREJO 2. CULTURA IN VITRO 3. MELHORAMENTO GENÉTICO VEGETAL 4. SENESCÊNCIA 5. ETAPA DE DESENVOLVIMENTO 6. CLOROFILA	1.- SENESCENCIA 2.- ACUCARES 3.- AREA FOLIAR 4.- CARVAO 5.- CLOROFILA 6.- ETILENO 7.- ENRAIZAMENTO	8.- INOCULACAO 9.- MICROPROPAGACAO 10.- MATERIA SECA 11.- SACAROSE 12.- SECA 13.- TRATAMENTO	$C_i = \frac{2}{(6+13)-2} = \frac{2}{17} = 0{,}11$
77	ZIETEMANN, Corina e ROBERTO, Sérgio Ruffo. Efeito de diferentes substratos e épocas de coleta no enraizamento de estacas herbáceas de goiabeira, cvs. paluma e século XXI. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 31-36. ISSN 0100-2945.	1. GOIABA 2. ESTACA 3. MATERIAL DE PROPAGAÇÃO 4. SUBSTRATO 5. PULVERIZAÇÃO 6. PROPAGAÇÃO VEGETATIVA	1.- ENRAIZAMENTO 2.- GOIABA 3.- ARROZ 4.- COMPRIMENTO 5.- CALO 6.- ESTACA 7.- IMERSAO	8.- PULVERIZACAO 9.- PRIMAVERA 10.- SECA 11.- SOBREVIVENCIA 12.- TRABALHO 13.- VERAO	$C_i = \frac{3}{(6+13)-3} = \frac{3}{16} = 0{,}18$
78	BOTELHO, Renato Vasconcelos e MULLER, Marcelo Marques Lopes. Extrato de alho como alternativa na quebra de dormência de gemas em macieiras cv. Fuji Kiku. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 37-41. ISSN 0100-2945.	1. MAÇÃ 2. BROTAÇÃO 3. QUEBRA DA DORMÊNCIA 4. GEMA 5. FLORAÇÃO 6. EXTRATO 7. ALHO 8. HIDROGENAÇÃO 9. PULVERIZAÇÃO 10. CITOQUININA 11. PARANÁ	1.- ALHO 2.- DORMENCIA 3.- EXTRATO 4.- BROTACAO 5.- FLORACAO 6.- FRUTICULTURA 7.- GEMA 8.- INVERNO 9.- MINERAL	10.- OLEO 11.- OLEO MINERAL 12.- PODA 13.- PRODUCAO 14.- PLANTA 15.- PRIMAVERA 16.- TRABALHO 17.- TRATAMENTO	$C_i = \frac{5{,}5}{(11+17)-5{,}5} = \frac{5{,}5}{22{,}5} = 0{,}24$

79	DANTAS, Bárbara França; RIBEIRO, Luciana de Sá e PEREIRA, Maiane Santos. Teor de açúcares solúveis e insolúveis em folhe de videiras, cv. syrah, em diferentes posições no ramo e épocas do ano. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 42-47. ISSN 0100-2945.	1. UVA 2. BIOQUÍMICA VEGETAL 3. AÇÚCAR 4. FOLHA 5. FENOLOGIA 6. CLIMA 7. TEMPERATURA 8. CARBOIDRATO 9. PERNAMBUCO	1.- ACUCARES 2.- RAMO 3.- CRESCIMENTO 4.- DEMANDA 5.- FOTOSSINTESE 6.- FISIOLOGIA 7.- FISIOLOGIA VEGETAL 8.- FOLHA 9.- INSOLACAO 10.- LABORATORIO 11.- MATURACAO	12.- PRODUCAO 13.- RADIACAO 14.- TRABALHO 15.- TEMPERATURA 16.- UVA 17.- VINHO 18.- VALE 19.- VARIACAO SAZONAL 20.- VITIS VINIFERA	$C_i = \frac{4}{(9+20)-4} = \frac{4}{25} = 0{,}16$
80	SANCHES, Juliana; DURIGAN, José Fernando e SANTOS, Jaime Maia dos. Utilização da microscopia eletrônica de varredura como ferramenta de avaliação da estrutura do tecido de abacate 'quintal' após danos mecânicos. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 57-60. ISSN 0100-2945.	1. ABACATE 2. DANO MECÂNICO 3. CORTE 4. MICROSCOPIA ELETRÔNICA 5. PROPRIEDADE ORGANOLÉPTICA 6. POLPA 7. TECIDO VEGETAL	1.- ABACATE 2.- FERRAMENTA 3.- MICROSCOPIA 4.- MICROSCOPIA ELETRONICA 5.- TECIDO 6.- ALTURA	7.- ARMAZENAMENTO 8.- CORTE 9.- COMPRIMENTO 10.- PESO 11.- PROFUNDIDADE	$C_i = \frac{3{,}5}{(7+11)-3{,}5} = \frac{3{,}5}{14{,}5} = 0{,}24$
81	GALHO, Adriana Silva; LOPES, Nei Fernandes; BACARIN, Marcos Antônio e LIMA, Maria da Graças de Souza. Composição química e respiração de crescimento em frutos de *Psidium cattleyanum* sabine durante o ciclo de desenvolvimento. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 61-66. ISSN 0100-2945.	1. ARAÇÁ 2. COMPOSIÇÃO QUÍMICA 3. GLICOSE 4. RESPIRAÇÃO 5. CARBOIDRATO 6. FISIOLOGIA 7. NUTRIENTE	1.- COMPOSICAO QUIMICA 2.- CRESCIMENTO 3.- QUIMICA 4.- RESPIRACAO 5.- ARACA 6.- AMIDO 7.- ACUCARES 8.- CAMPO 9.- CLIMA 10.- CUSTO	11.- DIA DE CAMPO 12.- ESTIMATIVA 13.- FRUTO 14.- GLICOSE 15.- IDADE 16.- MATURACAO 17.- POMAR 18.- PROTEINA 19.- TAXA	$C_i = \frac{4}{(7+19)-4} = \frac{4}{22} = 0{,}18$
82	SCANAVACA JUNIOR, Laerte; FONSECA, Nelson e PEREIRA, Mácio Eduardo Canto. Uso de fécula de mandioca na pó-colheita de manga 'surpresa'. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 67-71. ISSN 0100-2945.	1. MANGA 2. MANDIOCA 3. PÓS-COLHEITA 4. BIOFILME 5. EMBALAGEM 6. PRESERVAÇÃO DE ALIMENTO 7. AMIDO 8. VIDA-DE-PRATELEIRA 9. PERDA 10. ÁGUA 11. MÉTODO ESTATÍSTICO 12. BAHIA	1.- FECULA 2.- MANDIOCA 3.- MANGA 4.- POS-COLHEITA 5.- ACIDEZ 6.- AGUA 7.- AMIDO 8.- CASCA 9.- FRUTA	10.- FRUTO 11.- MASSA 12.- PERDA 13.- POLPA 14.- PH 15.- TEMPERATURA 16.- TRATAMENTO 17.- UMIDADE	$C_i = \frac{6}{(12+17)-6} = \frac{6}{23} = 0{,}26$

83	MONTEIRO, Lino Bittencourt et al. Avaliação de atrativos alimentares utilizados no monitoramento de mosca-das-frutas em pessegueiro na lapa- PR. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 72-74. ISSN 0100-2945.	1. PÊSSEGO 2. UVA 3. MOSCA DAS FRUTAS 4. CONTROLE BIOLÓGICO 5. SUCO 6. PARANÁ	1.- PESSEGO 2.- ANASTREPHA 3.- ATRATIVO 4.- PRODUCAO VEGETAL 5.- CAPTURA	6.- PROTEINA 7.- PELO 8.- SUCO 9.- UVA 10.- VINAGRE	$C_i = \frac{3}{(6+10)-3} = \frac{3}{13} = 0{,}23$
84	MARAFON, Anderson Carlos et al. Concentrações de carboidratos em tecidos de pessegueiro (*Prunus persica* (L.) Batsch) cv. jubileu em plantas com ou sem sintomas de morte-precoce durante o período de dormência. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 75-79. ISSN 0100-2945.	1. PÊSSEGO 2. TECIDO 3. GEMA 4. AMIDO 5. AÇÚCAR 6. QUEBRA DA DORMÊNCIA 7. FISIOLOGIA VEGETAL 8. CARBOIDRATO 9. COMPOSIÇÃO QUÍMICA 10. ANÁLISE QUÍMICA 11. SÍNDROME 12. SINTOMA 13. MORTE PRECOCE 14. RIO GRANDE DO SUL	1.- DORMENCIA 2.- PESSEGO 3.- AMIDO 4.- ACUCARES 5.- BROTACAO 6.- VARIEDADE 7.- FLORACAO	8.- FRUTOSE 9.- GLICOSE 10.- INVERNO 11.- PELO 12.- SINDROME	$C_i = \frac{3{,}5}{(14+12)-3{,}5} = \frac{3{,}5}{22{,}5} = 0{,}15$
85	SANTOS, Janaína Pereira dos et al. Parasitóides de lepidópteros minadores presentes em plantas de crescimento espontâneo em pomar orgânicos de citrus montenegro - RS. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 80-84. ISSN 0100-2945.	1. CITRICULTURA 2. LAGARTA MINADORA 3. LEPIDÓPTERO 4. CONTROLE BIOLÓGICO 5. PRAGA DE PLANTA 6. RIO GRANDE DO SUL	1.- CRESCIMENTO 2.- FRUTA CITRICA 3.- POMAR 4.- CONTROLE BIOLOGICO 5.- CITRICULTURA	6.- LABORATORIO 7.- PHYLLOCNISTIS CITRELLA 8.- TRABALHO	$C_i = \frac{2}{(6+8)-2} = \frac{2}{12} = 0{,}16$
86	SILVA, Marcos Zatti da e OLIVEIRA, Caralos Amadeu Leite de. Toxicidade residual de alguns agrotóxicos recomendado na agricultura sobre *Neoseiulus californicus* (McGregor) (Acari: Phytoseiidae). *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 85-90. ISSN 0100-2945.	1. FRUTA CÍTRICA 2. AGROTÓXICO 3. ÁCARO 4. CONTROLE BIOLÓGICO 5. TOXIDEZ 6. EFEITO RESIDUAL	1.- TOXIDEZ 2.- ALIMENTO 3.- ACARO 4.- FRUTA CITRICA 5.- CAMPO 6.- CONTROLE BIOLOGICO 7.- DIA DE CAMPO 8.- ENXOFRE 9.- IDADE 10.- LABORATORIO	11.- METODO 12.- MORTALIDADE 13.- OXIDO 14.- PERA 15.- PREDADOR 16.- SUBSIDIO 17.- TRABALHO 18.- TETRANYCHUS URTICAE 19.- VARIEDADE	$C_i = \frac{4}{(6+19)-4} = \frac{4}{21} = 0{,}19$
87	GUERRA, Denis Salvati et al. Utilização de pesticidas na produção de pêssegos 'Marli', nos sistemas de produção integrada e convencional. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 91-95. ISSN 0100-2945.	1. PÊSSEGO 2. PRAGA DE PLANTA 3. INSETO 4. PESTICIDA 5. RIO GRANDE DO SUL	1.- PRODUCAO 2.- CONDICAO AMBIENTAL 3.- INGREDIENTE 4.- INSETO 5.- PESSEGO	6.- RIO 7.- SAFRA	$C_i = \frac{2}{(5+7)-2} = \frac{2}{10} = 0{,}20$

88	LUCENA, Eliseu Marlônio Pereira de et al. Alterações físicas e químicas durante o desenvolvimento de mangas 'Tommy Atkins' no vale de São Francisco, Petrolina-PE. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 96-101. ISSN 0100-2945.	1. MANGA 2. PÓS-COLHEITA 3. PROPAGAÇAO VEGETATIVA 4. ANÁLISE QUÍMICA 5. ANÁLISE FÍSICA 6. IRRIGAÇÃO 7. FISIOLOGIA VEGETAL 8. PERNAMBUCO	1.- VALE 2.- AGUA 3.- CRESCIMENTO 4.- COLHEITA 5.- COR 6.- CASCA 7.- DIAMETRO 8.- FISIOLOGIA 9.- FRUTO	10.- MANGA 11.- MASSA 12.- MATURIDADE 13.- PRODUTO 14.- POS-COLHEITA 15.- SECA 16.- TRABALHO	$C_i = \frac{2,5}{(8+16)-2,5} = \frac{2,5}{21,5} = 0,11$
89	SILVA, José Tadeu Alves da; PACHECO, Dilermando Dourado e COSTA, Édio Luiz da. Atributos químicos e físicos de solos cultivados com bananeiras 'Prata-Anã' (AAB), em três níveis de produtividade, no norte de Minas Gerais. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 102-106. ISSN 0100-2945.	1. BANANA 2. EXIGÊNCIA NUTRICIONAL 3. SOLO 4. FÍSICA DO SOLO 5. QUÍMICA DO SOLO 6. CÁLCIO 7. POTÁSSIO 8. MAGNÉSIO 9. MINAS GERAIS	1.- BANANA 2.- PRODUTIVIDADE 3.- ARGILA 4.- BANANAL 5.- VARIEDADE	6.- CONDUTIVIDADE ELETRICA 7.- SOLO 8.- SATURACAO 9.- TRABALHO	$C_i = \frac{2}{(9+9)-2} = \frac{2}{16} = 0,12$
90	RUFATO, Leo et al. Coberturas vegetais no desenvolvimento vegetativo de plantas de pessegueiro. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 107-109. ISSN 0100-2945.	1. PÊSSEGO 2. CONSORCIAÇÃO DE CULTURA 3. PROPAGAÇAO VEGETATIVA 4. COBERTURA VEGETAL 5. ECOLOGIA VEGETAL 6. PLANTA DE COBERTURA 7. PLANTA FORRAGEIRA 8. RIO GRANDE DO SUL	1.- PESSEGO 2.- COBERTURA VEGETAL 3.- CHICHARO 4.- ERVILHA 5.- PLANTA FORRAGEIRA 6.- INVERNO 7.- MANEJO 8.- NABO	9.- COLZA 10.- POMAR 11.- RIO 12.- SOLO 13.- TRABALHO 14.- VEGETACAO	$C_i = \frac{3}{(8+14)-3} = \frac{3}{19} = 0,15$
91	EDER-SILVA, Erllens; FELIX, Leonardo Pessoa e BRUNO, Riselane de Lucena Alcântara. Citogenética de algumas espécies frutíferas nativas do nordeste do Brasil. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 110-114. ISSN 0100-2945.	1. ABACAXI 2. CAJÁ 3. FRUTA TROPICAL 4. CARIOLOGIA 5. CITOLOGIA 6. CROMOSSOMO 7. MORFOLOGIA 8. ANÁLISE BIOLÓGICA 9. BIOQUÍMICA VEGETAL 10. MELHORAMENTO GENÉTICO VEGETAL 11. NORDESTE	1.- CITOGENETICA 2.- ANALISE 3.- ACIDO 4.- ARACA 5.- BROMELIA 6.- CARIOTIPO 7.- ESPECIE 8.- FAMILIA	9.- GUAZUMA ULMIFOLIA 10.- MYRTACEAE 11.- MELHORAMENTO 12.- REGISTRO 13.- SAPINDACEAE 14.- TRABALHO	$C_i = \frac{1}{(11+14)-1} = \frac{1}{24} = 0,04$

92	ZACCARO, Ronaldo Posella; DONADIO, Luiz Carlos; LEMOS, Eliana Gertrudes Macedo e PERECIN, Dilermando. Comportamento de cultivares de manga (*Mangifera indica* L.) em relação à malformação. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 115-119. ISSN 0100-2945.	1. MANGA 2. DOENÇA DE PLANTA 3. FUNGO 4. VARIEDADE 5. PORTA ENXERTO 6. INOCULAÇÃO 7. NORDESTE	1.- MANGA 2.- MALFORMACAO 3.- AMORA 4.- VARIEDADE 5.- DOENCA	6.- FUNGO 7.- FUSARIUM 8.- INOCULACAO 9.- PELO 10.- PRODUCAO	$C_i = \frac{4,5}{(7+10)-4,5} = \frac{4,5}{12,5} = 0,36$
93	SANTOS, Karine Louise dos et al. Evidência da atuação do sistema de auto-incompatibilidade tardia em *Acca Sellowiana* (berg) burret. (Myrtaceae). *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 120-123. ISSN 0100-2945.	1. GOIABA SERRANA 2. DICOTILEDÔNEA 3. MYRTACEAE 4. REPRODUÇÃO VEGETAL 5. POLINIZAÇÃO 6. GENÓTIPO 7. MELHORAMENTO GENÉTICO VEGETAL 8. SUL	1.- CRESCIMENTO 2.- GERMINACAO 3.- OVARIO 4.- POLEN	5.- POLINIZACAO 6.- TUBO POLINICO 7.- TRATAMENTO	$C_i = \frac{1}{(8+7)-1} = \frac{1}{14} = 0,07$
94	BELLON, Graciele et al. Variabilidade genética de acessos silvestres e comerciais de *Passiflora edulis* Sims. com base em marcadores RAPD. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 124-127. ISSN 0100-2945.	1. MARACUJÁ 2. MELHORAMENTO GENÉTICO VEGETAL 3. COMPORTAMENTO DE VARIEDADE 4. MARCADOR MOLECULAR 5. VARIAÇÃO GENÉTICA 6. GENÓTIPO 7. CERRADO	1.- GENETICA 2.- PASSIFLORA EDULIS 3.- CERRADO 4.- DNA 5.- DISPERSAO 6.- MELHORAMENTO	7.- MARACUJA 8.- MATRIZ 9.- RESISTENCIA 10.- TRABALHO	$C_i = \frac{3}{(7+10)-3} = \frac{3}{14} = 0,21$
95	VIEIRA, Renato Luíz; LEITE, Gabriel Berenhauser e WAMSER, Anderson Fernando. Efeito de substratos porosos no enraizamento *in vitro* do porta-enxerto de macieira M-9 (*Malus pumilla*). *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 128-132. ISSN 0100-2945.	1. MAÇÃ 2. PORTA ENXERTO 3. CULTURA IN VITRO 4. SUBSTRATO DE CULTURA 5. ENRAIZAMENTO 6. MICROPROPAGAÇÃO 7. ACLIMATAÇÃO 8. SANTA CATARINA	1.- ENRAIZAMENTO 2.- MACA 3.- AGAR 4.- CINZA 5.- COMPRIMENTO 6.- CRESCIMENTO 7.- GRANULOMETRIA 8.- MICROPROPAGACAO	9.- SOBREVIVENCIA 10.- TRABALHO 11.- TRATAMENTO 12.- TEMPERATURA 13.- TAXA 14.- VERMICULITA 15.- VIDRO 16.- VEGETACAO	$C_i = \frac{3}{(8+16)-3} = \frac{3}{21} = 0,14$
96	PIO, Rafael et al. Emergência e desenvolvimento de plântulas de cultivares de marmeleiro para uso como porta-enxertos. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 133-136. ISSN 0100-2945.	1. MARMELO 2. PROPAGAÇÃO VEGETATIVA 3. PORTA ENXERTO 4. VARIEDADE 5. MELHORAMENTO GENÉTICO VEGETAL	1.- EMERGENCIA 2.- AGUA 3.- ALTURA 4.- AREIA 5.- CURRAL 6.- CYDONIA OBLONGA 7.- DIAMETRO 8.- ESTERCO 9.- FRIO 10.- MASSA	11.- PARTE AEREA 12.- PERFORMANCE 13.- SEMEADURA 14.- SECA 15.- SOLO 16.- TRABALHO 17.- VERMICULITA 18.- VIVEIRO	$C_i = \frac{0}{(5+18)-0} = \frac{0}{23} = 0$

97	ZIETEMANN, Corina e ROBERTO, Sérgio Ruffo. Produção de mudas de goiabeira (*Psidium guajava* L.) em diferentes substratos. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 137-142. ISSN 0100-2945.	1. GOIABA 2. MUDA 3. TRANSPLANTE DE PLANTA 4. ENXERTO 5. ESTACA 6. ENRAIZAMENTO DE ESTACA 7. SUBSTRATO DE CULTURA	1.- GOIABA 2.- PRODUCAO 3.- COCO 4.- COMPRIMENTO 5.- ENRAIZAMENTO 6.- FIBRA 7.- MISTURA	8.- PARTE AEREA 9.- PLANTA 10.- SOLO 11.- SECA 12.- TRABALHO 13.- TRANSPLANTE DE PLANTA	$C_i = \frac{2,5}{(7+13)-2,5} = \frac{2,5}{17,5} = 0,14$
98	TEIXEIRA, Luiz Antônio Junqueira; NATALE, William; BETTIOL NETO, José Emílio e MARTINS, Antonio Lúcio Mello. Nitrogênio e potássio em bananeira via fertirrigação e adubação convencional-atributos químicos do solo. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 143-152. ISSN 0100-2945.	1. BANANA 2. ANÁLISE DO SOLO 3. QUÍMICA DO SOLO 4. CLORETO DE POTÁSSIO 5. NITRATO DE AMÔNIO 6. FERTIRRIGAÇÃO 7. ADUBAÇÃO 8. ACIDEZ DO SOLO 9. RELAÇÃO SOLO-PLANTA 10. SÃO PAULO	1.- ADUBACAO 2.- BANANA 3.- FERTIRRIGACAO 4.- NITROGENIO 5.- POTASSIO 6.- SOLO 7.- ACIDEZ 8.- ADUBO 9.- ANALISE 10.- DOSE	11.- NITRATO 12.- NITRATO DE AMONIO 13.- PRODUCAO 14.- PERFIL DO SOLO 15.- PH 16.- PROFUNDIDADE 17.- SATURACAO 18.- SOLIDO	$C_i = \frac{5,5}{(10+18)-5,5} = \frac{5,5}{22,5} = 0,24$
99	TEIXEIRA, Luiz Antônio Junqueira; NATALE, William e MARTINS, Antônio Lúcio Mello. Nitrogênio e potássio via fertirrigação e adubação convencional-estado nutricional das bananeiras e produção de frutos. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 153-160. ISSN 0100-2945.	1. BANANA 2. ANÁLISE FOLIAR 3. CLORETO DE POTÁSSIO 4. NITRATO DE AMÔNIO 5. FERTIRRIGAÇÃO 6. ADUBAÇÃO 7. PROPAGAÇAO VEGETATIVA 8. MODELO MATEMÁTICO 9. SÃO PAULO	1.- ADUBACAO 2.- FERTIRRIGACAO 3.- NITROGENIO 4.- POTASSIO 5.- PRODUCAO 6.- ANALISE	7.- CRESCIMENTO 8.- DOSE 9.- ESTADO NUTRICIONAL 10.- REDUCAO	$C_i = \frac{3}{(9+10)-3} = \frac{3}{16} = 0,18$
100	COELHO, Ruimário Inácio et al. Resposta à adubação com uréia, cloreto de potássio e ácido bórico em mudas abacaxizeiro'Smooth Cayenne'. *Rev. Bras. Frutic.* [online]. 2007, vol.29, n.1, pp. 161-165. ISSN 0100-2945.	1. ABACAXI 2. ADUBAÇÃO FOLIAR 3. MUDA 4. RESPOSTA DA PLANTA 5. MÉTODO ESTATÍSTICO	1.- ADUBACAO 2.- ACIDO 3.- ACIDO BORICO 4.- ABACAXI 5.- POTASSIO 6.- UREIA 7.- ADUBACAO FOLIAR	8.- ALTURA 9.- AREA FOLIAR 10.- CAULE 11.- CRESCIMENTO 12.- MASSA 13.- PLANTIO 14.- SECA	$C_i = \frac{2}{(5+14)-2} = \frac{2}{17} = 0,11$
			ÍNDICE MÉDIO DE CONSISTÊNCIA NA INDEXAÇÃO		19, 22 x 100%= 1922% 1922% /100 artigos = 19,22% **Total: 19,22%**

APÊNDICE E - Assuntos de cada artigo científico

Artigos científicos	Relevantes para:	Artigos científicos	Relevantes para:
Artigo 1	PÊRA; CARBOIDRATO; GEMA; CLIMA TEMPERADO; AMIDO; INVERNO	Artigo 51	PÊSSEGO; PORTA ENXERTO; CLONE
Artigo 2	PESSEGO; PÓLEN; GERMINAÇÃO	Artigo 52	AGRONEGÓCIO; MAÇÃ; RENTABILIDADE; CUSTO DE PRODUÇÃO
Artigo 3	SEMENTE; TANGERINA; DESIDRATAÇÃO	Artigo 53	POLINIZAÇÃO; LARANJA; PRODUÇÃO DE SEMENTES
Artigo 4	LICHIA; FRUTIFICAÇÃO; MATURAÇÃO; FRUTO	Artigo 54	CONDIÇÃO AMBIENTAL; LARANJA; FLORAÇÃO; INDUÇÃO; FRUTA CÍTRICA
Artigo 5	LICHIA; ANELAGEM; FLORAÇÃO; FRUTIFICAÇÃO	Artigo 55	AGROTÓXICO; MAÇÃ; CONTROLE INTEGRADO
Artigo 6	MYRTACEAE; EUGENIA INVOLUCRATA; GERMINAÇÃO; PÓLEN	Artigo 56	VINHO; UVA; VARIEDADE RESISTENTE
Artigo 7	MARACUJÁ; MATURAÇÃO; PÓS-COLHEITA; SEMENTE; GERMINAÇÃO	Artigo 57	SACAROSE; CULTURA IN VITRO; MARACUJÁ
Artigo 8	MANGABA; SEMENTE; TESTE DE VIGOR; EXTRAÇÃO	Artigo 58	MANGA; MATURAÇÃO; FRUTO; ARMAZENAMENTO; EMBALAGEM
Artigo 9	PÊRA; BORO; CÁLCIO; GEMA	Artigo 59	MARACUJÁ; PROPAGAÇÃO VEGETATIVA; ESTACA; ENRAIZAMENTO
Artigo 10	PÊSSEGO; FLORAÇÃO; BROTAÇÃO; FRUTIFICAÇÃO	Artigo 60	CARVÃO; BANANA; PROPAGAÇÃO VEGETATIVA; CULTURA IN VITRO
Artigo 11	ARMAZENAMENTO; PITANGA; REFRIGERAÇÃO; PÓS-COLHEITA; MATURAÇÃO	Artigo 61	MARACUJÁ; ENRAIZAMENTO; VARIEDADE; ESTACA
Artigo 12	ARMAZENAMENTO; PITANGA; PÓS-COLHEITA; PRESERVAÇÃO DE ALIMENTO; MATURAÇÃO	Artigo 62	SEMENTE; GERMINAÇÃO; TEMPERATURA; PLANTULA
Artigo 13	MAÇÃ; PÓS-COLHEITA; DISTÚRBIO FISIOLÓGICO	Artigo 63	MICROPROPAGAÇÃO; PEQUI; BROTAÇÃO INDUZIDA
Artigo 14	PÓS-COLHEITA; MANGA; PRESERVAÇÃO DE ALIMENTO	Artigo 64	ENXERTO; PORTA ENXERTO; PÊSSEGO
Artigo 15	ADUBO VERDE; ADUBAÇÃO VERDE; LARANJA; FRUTA CÍTRICA; QUALIDADE	Artigo 65	MARACUJÁ; FRUTO; HIDROGÊNIO; FERTIRRIGAÇÃO
Artigo 16	MATURAÇÃO; CAQUI; FRUTO; PÓS-COLHEITA; QUALIDADE	Artigo 66	BORO; SOLO; MARACUJÁ; ADUBAÇÃO
Artigo 17	TANGOR MURCOTT; PORTA ENXERTO	Artigo 67	FERTIRRIGAÇÃO; POTÁSSIO; NITROGÊNIO; COCO; ADUBAÇÃO
Artigo 18	MAMÃO; PÓS-COLHEITA; FRUTO; PROPRIEDADE FÍSICO-QUÍMICA; FENÓTIPO	Artigo 68	BANANA; ANÁLISE FOLIAR; GENÓTIPO; NUTRIÇÃO VEGETAL
Artigo 19	ACEROLA; ESTACA; MUDA; REPRODUÇÃO VEGETAL	Artigo 69	BANANA; PÓS-COLHEITA; PRESERVAÇÃO DE ALIMENTO; FRUTO; QUALIDADE
Artigo 20	FRUTA CÍTRICA; COVA; LATOSSOLO; RAÍZ; SOLO	Artigo 70	MAMÃO; MATURAÇÃO; PÓS-COLHEITA; ETILENO; ARMAZENAMENTO
Artigo 21	MARACUJÁ; FLORAÇÃO; FRUTIFICAÇÃO; ILUMINAÇÃO ARTIFICIAL; IRRIGAÇÃO; ENTRESSAFRA	Artigo 71	LARANJA; BORO; FRUTO; QUALIDADE
Artigo 22	MARACUJÁ; CLONE; PORTA ENXERTO; PRODUTIVIDADE; REPRODUÇÃO VEGETAL; ESTACA; VARIEDADE RESISTENTE	Artigo 72	BANANA; MATURAÇÃO; PÓS-COLHEITA
Artigo 23	BANANA; VARIEDADE; PRODUTIVIDADE	Artigo 73	PODA; RALEIO; PRODUTIVIDADE; TANGERINA; FRUTO
Artigo 24	UVA; MELHORAMENTO GENÉTICO VEGETAL; MARCADOR MOLECULAR; FENÓTIPO	Artigo 74	ARATICUM; CERRADO; MEIO AMBIENTE
Artigo 25	BANANA; ADUBAÇÃO; ADUBO ORGÂNICO; POTÁSSIO; NUTRIÇÃO VEGETAL; ANÁLISE FOLIAR	Artigo 75	GOIABA; FENOLOGIA; PODA
Artigo 26	MARACUJÁ; COCO; NITROGÊNIO; POTÁSSIO; FERTIRRIGAÇÃO; ANÁLISE FOLIAR	Artigo 76	ANNONACEAE; SENESCENCIA
Artigo 27	JAMBO; SEMENTE; GERMINAÇÃO; MORFOLOGIA VEGETAL; FRUTO	Artigo 77	GOIABA; ENRAIZAMENTO DE ESTACA; SUBSTRATO; PROPAGAÇÃO VEGETATIVA
Artigo 28	ATEMÓIA; ABSORÇÃO DE ÁGUA; SEMENTE	Artigo 78	MAÇÃ; QUEBRA DA DORMÊNCIA; EXTRATO; ALHO; GEMA
Artigo 29	GERMINAÇÃO; MARACUJÁ; ENXERTO; PROPAGAÇÃO VEGETATIVA	Artigo 79	UVA; AÇUCARES; RAMO
Artigo 30	MAÇÃ; QUEBRA DA DORMÊNCIA; BROTAÇÃO; PRODUTO QUÍMICO	Artigo 80	MICROSCOPIA ELETRÔNICA; ABACATE; DANO MECÂNICO
Artigo 31	COLLETOTRICHUM GLOEOSPORIOIDES; FUNGO; MARACUJÁ; MANGA; MAMÃO; GOIABA; ANTRACNOSE	Artigo 81	ARAÇA; GLICOSE; COMPOSIÇÃO QUÍMICA

Artigo 32	MAMÃO; DOENÇA DE PLANTA; PHYTOPHTHORA PALMIVORA	Artigo 82	MANGA; PÓS-COLHEITA; PRESERVAÇÃO DE ALIMENTO; AMIDO; BIOFILME
Artigo 33	UVA; FERTILIDADE; GEMA	Artigo 83	PÊSSEGO; MOSCAS DAS FRUTAS; ATRATIVO
Artigo 34	BANANA; GENÓTIPO; VARIEDADE RESISTENTE	Artigo 84	PÊSSEGO; CARBOIDRATO; AMIDO; AÇUCARES; GEMA; DORMÊNCIA
Artigo 35	VARIEDADE RESISTENTE; CLONE; LARANJA; DOENÇA DE PLANTA; FRUTA CÍTRICA; BACTÉRIA	Artigo 85	FRUTA CÍTRICA; CONTROLE BIOLÓGICO; HOSPEDEIRO
Artigo 36	PROPAGAÇÃO VEGETATIVA; LARANJA; BORBULHIAS; ENXERTO	Artigo 86	TOXIDEZ; AGROTOXICO; FRUTA CÍTRICA; PREDADOR
Artigo 37	BANANA; ADUBAÇÃO; FERTILIZANTE NITROGENADO; FERTILIZANTE POTÁSSICO; PRODUTIVIDADE	Artigo 87	PÊSSEGO; PESTICIDA; SISTEMA DE PRODUÇÃO; DEFENSIVO
Artigo 38	ÁCIDO GIBERÉLICO; MARACUJÁ; FLORAÇÃO; INDUÇÃO	Artigo 88	MANGA; FRUTO; PÓS-COLHEITA; COLHEITA; ARMAZENAMENTO
Artigo 39	MAÇÃ; EMBALAGEM; FRUTO; CONTROLE BIOLÓGICO	Artigo 89	BANANA; SOLO; FÍSICA DO SOLO; QUIMICA DO SOLO; PRODUTIVIDADE
Artigo 40	FRUTA CÍTRICA; RAIZ; ENRAIZAMENTO; SISTEMA RADICULAR	Artigo 90	PÊSSEGO; COBERTURA VEGETAL
Artigo 41	PÓS-COLHEITA; CEREJA; PRESERVAÇÃO DE ALIMENTO	Artigo 91	CARIÓTIPO; ARVORE FRUTÍFERA; CITOGENÉTICA
Artigo 42	PÓS-COLHEITA; MORANGO; ARMAZENAMENTO; PRESERVAÇÃO DE ALIMENTO; QUALIDADE	Artigo 92	MANGA; FUNGO; MALFORMAÇÃO; VARIEDADE RESISTENTE
Artigo 43	PÓS-COLHEITA; MARACUJÁ; PRESERVAÇÃO DE ALIMENTO; CARNAÚBA; PLÁSTICO; ARMAZENAMENTO	Artigo 93	GOIABA SERRANA; REPRODUÇÃO VEGETAL; POLINIZAÇÃO; MYRTACEAE
Artigo 44	FRUTA CÍTRICA; PRAGA DE PLANTA; CONTROLE INTEGRADO; CONTROLE BIOLÓGICO	Artigo 94	MARACUJÁ; VARIAÇÃO GENÉTICA; MARCADOR MOLECULAR
Artigo 45	FRUTA CÍTRICA; XYLELLA FASTIDIOSA; BACTÉRIA; XILEMA	Artigo 95	SUBSTRATO; ENRAIZAMENTO; MAÇÃ; PORTA ENXERTO; PROPAGAÇÃO VEGETATIVA
Artigo 46	AGROTÓXICO; CITRICULTURA; FRUTA CÍTRICA; ÁCARO; CONTROLE BIOLÓGICO	Artigo 96	MARMELO; PORTA ENXERTO; ENXERTO
Artigo 47	PLANTA HOSPEDEIRA; ÁCARO; LEPROSE; ERVA DANINHA	Artigo 97	GOIABA; PROPAGAÇÃO VEGETATIVA; ENRAIZAMENTO; SUBSTRATO; MUDA
Artigo 48	CADEIA PRODUTIVA; NESPERA; MERCADO	Artigo 98	FERTIRRIGAÇÃO; ADUBAÇÃO; BANANA; NITROGÊNIO; POTÁSSIO; FERTILIZANTE POTÁSSICO; FERTILIZANTE NITROGENADO
Artigo 49	PÊSSEGO; COMERCIALIZAÇÃO; SABOR; PREÇO	Artigo 99	FERTIRRIGAÇÃO; BANANA; ADUBAÇÃO; FERTILIZANTE POTASSICO; FERTILIZANTE NITROGENADO; PRODUTIVIDADE; POTÁSSIO; NITROGENIO
Artigo 50	TANGERINA; VARIEDADE; PORTA ENXERTO	Artigo 100	ABACAXI; ADUBAÇÃO FOLIAR; UREIA; CLORETO DE POTASSIO; ACIDO BORICO

APÊNDICE F - Necessidades de informação e respectivos artigos científicos relevantes na base de dados

Necessidades de informação:	Estratégia de busca	Artigos relevantes nas bases de dados:
1. Artigos sobre Adubação de bananeiras	*Adubação* E *Banana*	Artigos 25; 37; 98 e 99
2. Artigos sobre Fertirrigação com potássio	*Fertirrigação* E *Potássio*	Artigos 26; 67; 98 e 99
3. Artigos sobre Adubação verde	*Adubação verde*	Artigo 15
4. Artigos sobre Análise foliar de bananeiras	*Análise foliar* E *Banana*	Artigos 25 e 68
5. Artigos sobre maturação em pós-colheita	*Maturação* E *Pós-colheita*	Artigos 7; 11; 12; 16; 70 e 72
6. Artigos sobre armazenamento de frutos em pós-colheita	*Armazenamento* E *Pós-colheita*	Artigos 11; 12; 42; 43; 70 e 88
7. Artigos sobre conservação de frutos em pós-colheita	*Preservação de alimento* E *Pós-colheita*	Artigos 12; 14; 41; 42; 43; 69 e 82
8. Artigos sobre pós-colheita de manga	*Pós-colheita* E *Manga*	Artigos 14; 82 e 88
9. Artigos sobre armazenamento de pitangas	*Armazenamento* E *Pitanga*	Artigos 11 e 12
10. Artigos sobre porta enxerto de pêssegos	*Porta enxerto* E *Pêssego*	Artigos 51 e 64
11. Artigos sobre enraizamento de maracujá	*Enraizamento* E *Maracujá*	Artigos 59 e 61
12. Artigos sobre propagação vegetativa de maracujá	*Propagação vegetativa* E *Maracujá*	Artigos 29 e 59
13. Artigos sobre germinação de Sementes	*Germinação* E *Semente*	Artigos 7; 27 e 62
14. Artigos sobre testes em sementes de mangaba	*Teste de vigor* E *Semente* E *Mangaba*	Artigos 8
15. Artigos sobre frutificação de lichia	*Frutificação* E *Lichia*	Artigos 4 e 5
16. Artigos sobre solo de plantio de citros	*Solo* E *Fruta cítrica*	Artigo 20
17. Artigos sobre floração induzida	*Indução* E *Floração*	Artigos 38 e 54
18. Artigos que associe poda à produção de frutos	*Poda* E *Produtividade*	Artigo 73
19. Artigos sobre produtividade de bananas	*Produtividade* E *Banana*	Artigos 23; 37; 89 e 99
20. Artigos sobre quebra da dormência em macieiras	*Quebra da dormência* E *Maçã*	Artigos 30 e 78
21. Artigos sobre variedades resistentes de bananeiras	*Variedade resistente* E *Banana*	Artigo 34
22. Artigos sobre controle biológico de frutas cítricas	*Controle biológico* E *Fruta cítrica*	Artigos 44; 46 e 85
23. Artigos sobre fungo em mangas	*Fungo* E *Manga*	Artigos 31 e 92
24. Artigos sobre enraizamento de goiabeiras	*Enraizamento* E *Goiaba*	Artigos 77 e 97
25. Artigos sobre uso de agrotóxicos em frutas cítricas	*Agrotóxico* E *Fruta cítrica*	Artigos 46 e 86
26. Artigos sobre uso de biofilme em pós-colheita de manga	*Biofilme* E *Pós-colheita* E *Manga*	Artigo 82
27. Artigos sobre frutificação de maracujá	*Frutificação* E *Maracujá*	Artigos 21
28. Artigos sobre sementes de atemóia	*Semente* E *Atemóia*	Artigo 28
29. Artigos sobre ácido giberélico na indução de floração	*Ácido giberélico* E *Floração*	Artigo 38
30. Artigos sobre variedades resistentes de videiras	*Variedade resistente* E *Uva*	Artigo 56
31. Artigos sobre adubação por irrigação	*Fertirrigação*	Artigos 26; 65; 67; 98 e 99
32. Artigos sobre uso de uréia em plantações de abacaxi	*Uréia* E *Abacaxi*	Artigo 100
33. Artigos sobre maçã	*Maçã*	Artigos 13; 30; 39; 52; 55; 78 e 95
34. Artigos sobre propagação vegetativa	*Propagação vegetativa*	Artigos 29; 36; 59; 60; 77; 95 e 97
35. Artigos sobre conservação de frutos	*Preservação de alimento*	Artigos 12; 14; 41; 42; 43; 69 e 82
36. Artigos sobre micropropagação	*Micropropagação*	Artigo 63
37. Artigos sobre doenças de plantas	*Doença de planta*	Artigos 32 e 35

38.	Artigos sobre comercialização de pêssegos	*Comercialização* E *Pêssego*	Artigo 49
39.	Artigos sobre associação de técnica de anelagem e frutificação	*Anelagem* E *Frutificação*	Artigo 5
40.	Artigos sobre gemas de pereiras	*Gema* E *Pêra*	Artigos 1 e 9
41.	Artigos sobre maturação de mamão	*Maturação* E *Mamão*	Artigo 70
42.	Artigos sobre semente de jambo	*Semente* E *Jambo*	Artigo 27
43.	Artigos sobre análise foliar	*Análise foliar*	Artigos 25; 26 e 68
44.	Artigos sobre o mercado da nêspera	*Mercado* E *Nêspera*	Artigo 48
45.	Artigos sobre uso de substrato no enraizamento	*Substrato* E *Enraizamento*	Artigos 77; 95 e 97
46.	Artigos sobre bactéria de fruta cítrica	*Bactéria* E *Fruta cítrica*	Artigos 35 e 45
47.	Artigos sobre pós-colheita	*Pós-colheita*	Artigos 7; 11; 12; 13; 14; 16; 18; 41; 42; 43; 69; 70; 72; 82 e 88
48.	Artigos sobre reprodução de mudas de acerola por estaca	*Estaca* E *Acerola*	Artigo 19
49.	Artigos sobre melhoramento genético de uvas	*Melhoramento genético vegetal* E *Uva*	Artigo 24
50.	Artigos sobre pólen de pêssego	*Pólen* E *Pêssego*	Artigo 2

APÊNDICE G - Cálculos de exaustividade e precisão na recuperação de informação em bases de dados BDSISA e BDBINAGRI

Base de dados BDSISA (Indexação A)	Base de dados BDBINAGRI (Indexação B)	Base de dados BDSISA (Indexação A)	Base de dados BDBINAGRI (Indexação B)
1ª Busca: *Adubação* E *Banana* Artigos relevantes: 25; 37; 98 e 99 Recuperados: 25; 37 e 98 Exaustividade = 3/4 = 0,75 =75% Precisão = 3/3 = 1= 100%	1ª Busca: *Adubação* E *Banana* Artigos relevantes: 25; 37; 98 e 99 Recuperados: 37; 98 e 99 Exaustividade = 3/4 = 0,75 = 75% Precisão = 3/3= 1 = 100%	2ª Busca: *Fertirrigação* E *Potássio* Artigos relevantes: 26; 67; 98 e 99 Recuperados: 26; 67; 98 e 99 Exaustividade = 4/4=1 =100% Precisão = 4/4 =1 =100%	2ª Busca: *Fertirrigação* E *Potássio* Artigos relevantes: 26; 67; 98 e 99 Recuperados: 26; 98 e 99 Exaustividade = 3/4= 0,75 = 75% Precisão = 3/3 = 1 = 100%
3ª Busca: *Adubação verde* Artigos relevantes: 15 Recuperados: 15 Exaustividade = 1/1=1=100% Precisão = 1/1=1= 100%	3ª Busca: *Adubação verde* Artigos relevantes: 15 Recuperados: 15 Exaustividade = 1/1=1= 100% Precisão = 1/1=1= 100%	4ª Busca: *Análise foliar* E *Banana* Artigos relevantes: 25 e 68 Recuperados: 68 Exaustividade = 1/2=0,50 = 50% Precisão =1/1= 1 = 100%	4ª Busca: *Análise foliar* E *Banana* Artigos relevantes: 25 e 68 Recuperados: 25; 68 e 99 Exaustividade = 2/2= 1= 100% Precisão =2/3= 0,66 = 66%
5ª Busca: *Maturação* E *Pós-colheita* Artigos relevantes: 7; 11; 12; 16; 70 e 72 Recuperados: 7; 11; 12; 13; 14; 16; 41; 69; 70 e 72 Exaustividade = 6/6= 1=100% Precisão =6/10=0,6 = 60%	5ª Busca: *Maturação* E *Pós-colheita* Artigos relevantes: 7; 11; 12; 16; 70 e 72 Recuperados: 7; 12; 14; 16 e 70 Exaustividade = 4/6= 0,66= 66% Precisão =4/5= 0,8= 80%	6ª Busca: *Armazenamento* E *Pós-colheita* Artigos relevantes: 11; 12; 42; 43; 70 e 88 Recuperados: 7; 11; 12; 14; 41; 43; 69 e 70 Exaustividade = 4/6=0,66 = 66% Precisão = 4/8= 0,5 = 50%	6ª Busca: *Armazenamento* E *Pós-colheita* Artigos relevantes: 11; 12; 42; 43; 70 e 88 Recuperados: 7; 11; 12; 14; 16; 41; 43 e 70 Exaustividade = 4/6= 0,66 = 66% Precisão = 4/8= 0,5 = 50%
7ª Busca: *Preservação de alimento* E *Pós-colheita* Artigos relevantes: 12; 14; 41; 42; 43; 69 e 82 Recuperados: 0 Exaustividade = 0/7= 0% Precisão =0/0= 0%	7ª Busca: *Preservação de alimento* E *Pós-colheita* Artigos relevantes: 12; 14; 41; 42; 43; 69 e 82 Recuperados: 12; 41 e 82 Exaustividade = 3/7= 0,42= 42% Precisão =3/3= 100%	8ª Busca: *Pós-colheita* E *Manga* Artigos relevantes: 14; 82 e 88 Recuperados: 14; 82 e 88 Exaustividade = 3/3= 1= 100% Precisão = 3/3= 1= 100%	8ª Busca: *Pós-colheita* E *Manga* Artigos relevantes: 14; 82 e 88 Recuperados: 14; 82 e 88 Exaustividade = 3/3= 1= 100% Precisão = 3/3= 1= 100%
9ª Busca: *Armazenamento* E *Pitanga* Artigos relevantes: 11 e 12 Recuperados: 11 Exaustividade = 1/2=0,5 = 50% Precisão =1/1= 1= 100%	9ª Busca: *Armazenamento* E *Pitanga* Artigos relevantes: 11 e 12 Recuperados: 11 e 12 Exaustividade = 2/2=1= 100% Precisão =2/2= 1= 100%	10ª Busca: *Porta enxerto* E *Pêssego* Artigos relevantes: 51 e 64 Recuperados: 0 Exaustividade = 0/2 =0% Precisão =0/0= 0%	10ª Busca: *Porta enxerto* E *Pêssego* Artigos relevantes: 51 e 64 Recuperados: 51 Exaustividade = 1/2 =0,5= 50% Precisão =1/1= 1= 100%
11ª Busca: *Enraizamento* E *Maracujá* Artigos relevantes: 59 e 61 Recuperados: 59 e 61 Exaustividade = 2/2= 1 = 100%	11ª Busca: *Enraizamento* E *Maracujá* Artigos relevantes: 59 e 61 Recuperados: 59 e 61 Exaustividade = 2/2= 1= 100%	12ª Busca: *Propagação vegetativa* E *Maracujá* Artigos relevantes: 29 e 59 Recuperados: 29 Exaustividade = 1/2 = 0,5= 50%	12ª Busca: *Propagação vegetativa* E *Maracujá* Artigos relevantes: 29 e 59 Recuperados: 29; 59 e 61 Exaustividade = 2/2 =1 = 100%

Precisão = 2/2= 1= 100%	Precisão = 2/2 = 1= 100%	Precisão =1/1= 1= 100%	Precisão =2/3= 0,66 = 66%
13ª Busca: *Germinação* E *Semente* Artigos relevantes: 7; 27 e 62 Recuperados: 3; 8 e 27 Exaustividade = 1/3= 0,33= 33% Precisão = 1/3= 0,33= 33%	13ª Busca: *Germinação* E *Semente* Artigos relevantes: 7; 27 e 62 Recuperados: 27; 28 e 62 Exaustividade = 2/3= 0,66= 66% Precisão = 2/3= 0,66= 66%	14ª Busca: *Teste de vigor* E *Semente* E *Mangaba* Artigos relevantes: 8 Recuperados: 0 Exaustividade = 0/1 = 0% Precisão =0/0= 0%	14ª Busca: *Teste de vigor* E *Semente* E *Mangaba* Artigos relevantes: 8 Recuperados: 8 Exaustividade = 1/1 = 100% Precisão =1/1= 100%
15ª Busca: *Frutificação* E *Lichia* Artigos relevantes: 4 e 5 Recuperados: 0 Exaustividade = 0/2 =0% Precisão =0/0= 0%	15ª Busca: *Frutificação* E *Lichia* Artigos relevantes: 4 e 5 Recuperados: 4 e 5 Exaustividade = 2/2 = 1= 100% Precisão =2/2= 1= 100%	16ª Busca: *Solo* E *Fruta cítrica*. Artigos relevantes: 20 Recuperados: 20 Exaustividade = 1/1= 1 = 100% Precisão =1/1= 1= 100%	16ª Busca: *Solo* E *Fruta cítrica*. Artigos relevantes: 20 Recuperados: 20 Exaustividade = 1/1= 100% Precisão =1/1= 100%
17ª Busca: *Indução* E *Floração* Artigos relevantes: 38 e 54 Recuperados: 38 e 54 Exaustividade = 2/2 =1 =100% Precisão =2/2=1 = 100%	17ª Busca: *Indução* E *Floração* Artigos relevantes: 38 e 54 Recuperados: 54 Exaustividade = 1/2 =0,5= 50% Precisão =1/1= 1 = 100%	18ª Busca: *Poda* E *Produtividade* Artigos relevantes: 73 Recuperados: 0 Exaustividade = 0/1= 0% Precisão =0/0= 0%	18ª Busca: *Poda* E *Produtividade* Artigos relevantes: 73 Recuperados: 0 Exaustividade = 0/1=0= 0% Precisão =0/0= 0%
19ª Busca: *Produtividade* E *Banana* Artigos relevantes: 23; 37; 89 e 99 Recuperados: 34; 37 e 89 Exaustividade = 2/4= 0,5 = 50% Precisão = 2/3= 0,66= 66%	19ª Busca: *Produtividade* E *Banana* Artigos relevantes: 23; 37; 89 e 99 Recuperados: 23 e 37 Exaustividade = 2/4= 0,5 = 50% Precisão = 2/2= 1= 100%	20ª Busca: *Quebra da dormência* E *Maçã* Artigos relevantes: 30 e 78 Recuperados: 0 Exaustividade = 0/2= 0% Precisão = 0/0= 0%	20ª Busca: *Quebra da dormência* E *Maçã* Artigos relevantes: 30 e 78 Recuperados: 30 e 78 Exaustividade 2/2=1 = 100% Precisão = 2/2= 1= 100%
21ª Busca: *Variedade resistente* E *Banana* Artigos relevantes: 34 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	21ª Busca: *Variedade resistente* E *Banana* Artigos relevantes: 34 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	22ª Busca: *Controle biológico* E *Fruta cítrica* Artigos relevantes: 44; 46 e 85 Recuperados: 46; 85 e 86 Exaustividade = 2/3= 0,66 = 66% Precisão = 2/3= 0,66 = 66%	22ª Busca: *Controle biológico* E *Fruta cítrica* Artigos relevantes: 44; 46 e 85 Recuperados: 46 e 86 Exaustividade = 1/3= 0,33 =33% Precisão = 1/2= 0,5 =50%
23ª Busca: *Fungo* E *Manga* Artigos relevantes: 31 e 92 Recuperados: 31 e 92 Exaustividade = 2/2= 1= 100% Precisão = 2/2= 1= 100%	23ª Busca: *Fungo* E *Manga* Artigos relevantes: 31 e 92 Recuperados: 31 e 92 Exaustividade = 2/2=1 = 100% Precisão = 2/2= 1= 100%	24ª Busca: *Enraizamento* E *Goiaba* Artigos relevantes: 77 e 97 Recuperados: 77 e 97 Exaustividade = 2/2= 1= 100% Precisão = 2/2= 1= 100%	24ª Busca: *Enraizamento* E *Goiaba* Artigos relevantes: 77 e 97 Recuperados: 97 Exaustividade = 1/2= 0,5= 50% Precisão = 1/1= 1= 100%
25ª Busca: *Agrotóxico* E *Fruta cítrica*	25ª Busca: *Agrotóxico* E *Fruta cítrica*	26ª Busca: *Biofilme* E *Pós-colheita* E *Manga*	26ª Busca: *Biofilme* E *Pós-colheita* E *Manga*

Artigos relevantes: 46 e 86 Recuperados: 0 Exaustividade = 0/2= 0% Precisão = 0/0= 0%	Artigos relevantes: 46 e 86 Recuperados: 86 Exaustividade = 1/2= 0,5= 50% Precisão = 1/1= 1= 100%	Artigos relevantes: 82 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	Artigos relevantes: 82 Recuperados: 82 Exaustividade = 1/1= 1= 100% Precisão = 1/1= 1 = 100%
27ª Busca: *Frutificação* E *Maracujá* Artigos relevantes: 21 Recuperados: 21 Exaustividade = 1/1=1= 100% Precisão = 1/1= 1= 100%	27ª Busca: *Frutificação* E *Maracujá* Artigos relevantes: 21 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	28ª Busca: *Semente* E *Atemóia* Artigos relevantes: 28 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	28ª Busca: *Semente* E *Atemóia* Artigos relevantes: 28 Recuperados: 28 Exaustividade = 1/1= 1= 100% Precisão = 1/1= 1= 100%
29ª Busca: *Ácido giberélico* E *Floração* Artigos relevantes: 38 Recuperados: 38 e 73 Exaustividade = 1/1= 1= 100% Precisão = 1/2= 0,5 = 50%	29ª Busca: *Ácido giberélico* E *Floração* Artigos relevantes: 38 Recuperados: 38 Exaustividade = 1/1= 1= 100% Precisão = 1/1= 1 = 100%	30ª Busca: *Variedade resistente* E *Uva* Artigos relevantes: 56 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	30ª Busca: *Variedade resistente* E *Uva* Artigos relevantes: 56 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%
31ª Busca: *Fertirrigação* Artigos relevantes: 26; 65; 67; 98 e 99 Recuperados: 26; 65; 67; 98 e 99 Exaustividade = 5/5=1= 100% Precisão = 5/5= 1= 100%	31ª Busca: *Fertirrigação* Artigos relevantes: 26; 65; 67; 98 e 99 Recuperados: 26; 65; 98 e 99 Exaustividade = 4/5= 0,8 = 80% Precisão = 4/4= 1= 100%	32ª Busca: *Uréia* E *Abacaxi* Artigos relevantes: 100 Recuperados: 100 Exaustividade = 1/1= 1= 100% Precisão = 1/1= 1 = 100%	32ª Busca: *Uréia* E *Abacaxi* Artigos relevantes: 100 Recuperados: 0 Exaustividade = 0/1= 0= 0% Precisão = 0/0= 0%
33ª Busca: *Maçã* Artigos relevantes: 13; 30; 39; 52; 55; 78 e 95 Recuperados: 30; 39; 52; 55 e 95 Exaustividade = 5/7= 0,71 = 71% Precisão = 5/5= 1= 100%	33ª Busca: *Maçã* Artigos relevantes: 13; 30; 39; 52; 55; 78 e 95 Recuperados: 13; 30; 39; 52; 55; 69; 78 e 95 Exaustividade = 8/8= 1 =100% Precisão = 8/8= 1= 100%	34ª Busca: *Propagação vegetativa* Artigos relevantes: 29; 36; 59; 60; 77; 95 e 97 Recuperados: 29 Exaustividade = 1/7=0,14= 14% Precisão = 1/1= 1= 100%	34ª Busca: *Propagação vegetativa* Artigos relevantes: 29; 36; 59; 60; 77; 95 e 97 Recuperados: 29; 36; 40; 51; 59; 60; 61; 63; 64; 77; 88; 90; 96 e 99 Exaustividade = 5/7= 0,71= 71% Precisão = 5/14= 0,35= 35%
35ª Busca: *Preservação de alimento* Artigos relevantes: 12; 14; 41; 42; 43; 69 e 82 Recuperados: 0 Exaustividade = 0/7= 0% Precisão = 0/0= 0%	35ª Busca: *Preservação de alimento* Artigos relevantes: 12; 14; 41; 42; 43; 69 e 82 Recuperados: 12; 41 e 82 Exaustividade = 3/7= 0,42 =42% Precisão = 3/3= 1= 100%	36ª Busca: *Micropropagação* Artigos relevantes: 63 Recuperados: 60; 63; 76 e 95 Exaustividade = 1/1=1=100% Precisão = 1/4= 0,25= 25%	36ª Busca: *Micropropagação* Artigos relevantes: 63 Recuperados: 95 Exaustividade = 0/1= 0% Precisão = 0/1= 0%
37ª Busca: *Doença de planta* Artigos relevantes: 32 e 35 Recuperados: 0 Exaustividade = 0/2= 0%	37ª Busca: *Doença de planta* Artigos relevantes: 32 e 35 Recuperados: 31; 35 e 92 Exaustividade = 1/2= 0,5 =50%	38ª Busca: *Comercialização* E *Pêssego* Artigos relevantes: 49 Recuperados: 49 Exaustividade = 1/1 =1 =100%	38ª Busca: *Comercialização* E *Pêssego* Artigos relevantes: 49 Recuperados: 49 Exaustividade = 1/1=1 = 100%

Precisão = 0/0= 0%	Precisão = 1/3= 0,33= 33%	Precisão = 1/1= 1= 100%	Precisão = 1/1= 1= 100%
39ª Busca: *Anelagem* E *Frutificação* Artigos relevantes: 5 Recuperados: 5 Exaustividade = 1/1=1 = 100% Precisão = 1/1= 1= 100%	39ª Busca: *Anelagem* E *Frutificação* Artigos relevantes: 5 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	40ª Busca: *Gema* E *Pêra* Artigos relevantes: 1 e 9 Recuperados: 0 Exaustividade = 0/2= 0% Precisão = 0/0= 0%	40ª Busca: *Gema* E *Pêra* Artigos relevantes: 1 e 9 Recuperados: 1 e 9 Exaustividade = 2/2= 1= 100% Precisão = 2/2= 1= 100%
41ª Busca: *Maturação* E *Mamão* Artigos relevantes: 70 Recuperados: 70 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%	41ª Busca: *Maturação* E *Mamão* Artigos relevantes: 70 Recuperados: 70 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%	42ª Busca: *Semente* E *Jambo* Artigos relevantes: 27 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	42ª Busca: *Semente* E *Jambo* Artigos relevantes: 27 Recuperados: 27 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%
43ª Busca: *Análise foliar* Artigos relevantes: 25; 26 e 68 Recuperados: 68 Exaustividade = 1/3= 0,33 = 33% Precisão = 1/1= 1= 100%	43ª Busca: *Análise foliar* Artigos relevantes: 25; 26 e 68 Recuperados: 9; 25; 26; 68 e 99 Exaustividade = 3/3= 1 =100% Precisão = 3/5= 0,6= 60%	44ª Busca: *Mercado* E *Nêspera* Artigos relevantes: 48 Recuperados: 48 Exaustividade = 1/1= 1= 100% Precisão = 1/1= 1= 100%	44ª Busca: *Mercado* E *Nêspera* Artigos relevantes: 48 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%
45ª Busca: *Substrato* E *Enraizamento* Artigos relevantes: 77; 95 e 97 Recuperados: 0 Exaustividade = 0/3= 0% Precisão = 0/0= 0%	45ª Busca: *Substrato* E *Enraizamento* Artigos relevantes: 77; 95 e 97 Recuperados: 59; 95 e 97 Exaustividade = 2/3= 0,66 = 66% Precisão = 2/3= 0,66= 66%	46ª Busca: *Bactéria* E *Fruta cítrica* Artigos relevantes: 35 e 45 Recuperados: 35 e 45 Exaustividade = 2/2= 1= 100% Precisão = 2/2= 1= 100%	46ª Busca: *Bactéria* E *Fruta cítrica* Artigos relevantes: 35 e 45 Recuperados: 0 Exaustividade = 0/2= 0% Precisão = 0/0= 0%
47ª Busca: *Pós-colheita* Artigos relevantes: 7; 11; 12; 13; 14; 16; 18; 41; 42; 43; 69; 70; 72; 82 e 88 Recuperados: 7; 11; 12; 13; 14; 16; 18; 41; 43; 69; 70; 72; 82 e 88 Exaustividade = 14/15= 0,93 =93% Precisão = 14/14= 1= 100%	47ª Busca: *Pós-colheita* Artigos relevantes: 7; 11; 12; 13; 14; 16; 18; 41; 42; 43; 69; 70; 72; 82 e 88 Recuperados: 7; 11; 12; 13; 14; 16; 18; 41; 43; 70; 72; 82 e 88 Exaustividade = 13/15= 0,86 =86% Precisão = 13/13= 1= 100%	48ª Busca: *Estaca* E *Acerola* Artigos relevantes: 19 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	48ª Busca: *Estaca* E *Acerola* Artigos relevantes: 19 Recuperados: 19 Exaustividade = 1/1=1 = 100% Precisão = 1/1= 1= 100%
49ª Busca: *Melhoramento genético vegetal* E *Uva* Artigos relevantes: 24 Recuperados: 0 Exaustividade = 0/1= 0% Precisão = 0/0= 0%	49ª Busca: *Melhoramento genético vegetal* E *Uva* Artigos relevantes: 24 Recuperados: 24 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%	50ª Busca: *Pólen* E *Pêssego* Artigos relevantes: 2 Recuperados: 2 Exaustividade = 1/1= 1 = 100% Precisão = 1/1= 1= 100%	50ª Busca: *Pólen* E *Pêssego* Artigos relevantes: 2 Recuperados: 2 Exaustividade = 1/1= 1 =100% Precisão = 1/1= 1= 100%