UNIVERSIDADE DE BRASÍLIA

Faculdade de Ciência da Informação

INDEXAÇÃO AUTOMÁTICA : UMA REVISÃO DE LITERATURA

Jainne Aragão Carvalho Fernandes

Orientadora: Profa. Dra. Simone Bastos Vieira

Brasília

2013

JAINNE ARAGÃO CARVALHO FERNANDES

INDEXAÇÃO AUTOMÁTICA: UMA REVISÃO DE LITERATURA

Monografia apresentada como parte das exigências para obtenção do título de Bacharel em Biblioteconomia pela Faculdade de Ciência da Informação da Universidade de Brasília

Orientadora: Profa. Dra. Simone Bastos Vieira

Brasília

2013

J758i

Fernandes, Jainne Aragão Carvalho
Indexação automática: uma revisão / Jainne Aragão Carvalho Fernandes. – 2013
98 f.

Monografia (Bacharelado em Biblioteconomia) – Universidade de Brasília, Faculdade de Ciência da Informação, Curso de Biblioteconomia, Brasília, 2013.

Orientação: Prof.ª Drª Simone Bastos Vieira.

1. Indexação automática. 2. Processamento de linguagem natural
I. Título

CDU 025

Titulo: Indexação automática: uma revisão de literatura

Aluna: Jainne Aragão Carvalho Fernandes

Monografia apresentada à Faculdade de Ciência da Informação da Universidade de Brasília, como parte dos requisitos para obtenção do grau de Bacharel em Biblioteconomia.

Brasília, 13 de dezembro de 2013.

Simone Bastos Vieira - Orientadora
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

Marcílio de Brito – Membro
Professor da Faculdade de Ciência da Informação (UnB)
Doutor em Ciências da Informação e da Documentação

Rita de Cássia do Vale Caribé – Membro
Professora da Faculdade de Ciência da Informação (UnB)
Doutora em Ciência da Informação

DEDICATÓRIA

Dedico o presente trabalho a meus pais e amigos que me apoiaram em todo o processo de construção deste estudo.

AGRADECIMENTOS

Agradeço primeiramente à Deus, por me dar a fé que precisava para continuar o caminho mesmo nos momentos mais turbulentos do processo.

Agradeço aos meus pais, pelo incentivo aos estudos, pelo apoio moral, e por todo amor, compreensão e carinho dedicados a mim em todos os momentos da vida.

Agradecimento especial à professora Simone Bastos, pelo auxílio na construção do trabalho, pela paciência, pela cobrança, e por estar sempre presente no decorrer de todo o processo.

Agradeço também aos meus amigos, que me deram suporte em vários momentos difícieis, me ajudando sempre a manter o foco.

Agradeço em especial ao meu chefe de estágio, Raphael Cavalcante, pelo apoio, pelas dicas de construção de texto e pela paciência.

Para o homem inteligente, o que lhe falta é justamente aquilo que o estimula.

(Paulo Coelho)

RESUMO

Trata-se de uma revisão de literatura acerca da indexação automática, por meio de pesquisas realizadas em bases de dados nacionais e internacionais. Procurou-se observar quantas publicações brasileiras e estrangeiras foram encontradas e quais foram as suas datas de publicação.

Foi mencionada a história da indexação de modo geral, e o surgimento da indexação automática como meio de recuperação da informação. Algumas problemáticas relacionadas com o presente tema foram colocadas em questão, como a sintaxe e a semântica no processo de indexação.

Ao final, mostrou-se também a relação da indexação nas bibliotecas digitais e na web; foram abarcados conceitos de web semântica e folksonomia, visando a relação de ambas com a indexação automática. Apresentou-se uma análise dos dados coletados ao longo do trabalho abarcando os últimos 32 anos de publicações em indexação automática. Além disso, foi realizada uma comparação com os estudos realizados por Ladeira (2010).

ABSTRACT

This work presents a literature review of automatic indexing through research in national and international data . The work tried to see how many Brazilian and foreign publications were found and what your dates of publication were.

It was mentioned the history of indexing in general, and the emergence of automatic indexing as a means of information retrieval. Some issues related to this theme were called into question, as the syntax and semantics in the indexing process .

At the end, also proved the relation of indexing in digital libraries and on the web ; concepts of semantic web and folksonomies were embraced , targeting the relationship with automatic indexing . The work also presented an analysis of data collected throughout the study covering the last 32 years of publications in automatic indexing. Furthermore, a comparison with the studies by Ladeira (2010) was performed .

LISTA DE SIGLAS

ARIST – Annual Review of Information Science and Technology

ASL – Análise Semântica Latente

EAGLES – Expert Advisory Group on Language Engineering Standards

EMR – Eletronic Medical Records

KWAC – Key-word-and-context

KWIC – Key-word-in-context

KWOC – Key-word-out-of-context

LIWC – Linguistic Inquiry and Word Count

MUC – Message Undersanding Conferences

PLN – Processamento de Linguagem Natural

RDF – Resource Description Framework

SRI – Sistema de Recuperação da Informação

URI – Uniform Resource Identifiers

LISTA DE GRÁFICOS E TABELAS

Sumário

1 INTRODUÇÃO

Em meio a grande explosão da informação, surge a necessidade de técnicas mais precisas de tratamento e recuperação da informação, uma vez que as técnicas manuais não seriam capazes de atender a grande demanda documental. Desse modo, a indexação automática surge como um fator primordial para uma recuperação eficiente tanto em ambientes físicos como em meios virtuais de acesso.

O presente trabalho é uma revisão de literatura dividida em tópicos. O primeiro apresentará o histórico da indexação de modo geral, seguido de suas definições básicas, em seguida trata da indexação automática, como surgiu e quais são suas tipologias; mais a frente serão brevemente mencionados os conceitos de linguística computacional.

Devido a sua importância para a indexação automática, o processamento da linguagem natural também é mencionado durante o trabalho, ressaltando-se a questão das ambiguidades e as aplicações do Processamento da Linguagem Natural (PLN) na documentação.

Seguindo na área de linguagem natural, apresenta-se a sintaxe e a semântica e suas relações com a indexação automática; aborda-se ainda os sintagmas nominais no contexto da recuperação da informação. O trabalho continua com uma sessão para apresentação dos experimentos em linguagem natural, onde serão sucintamente colocados experimentos brasileiros e estrangeiros que fazem uso de linguagem natural para uma indexação automática.

Em um último tópico são apresentadas as aplicações do PLN na internet, web e bibliotecas digitais. Por fim, é abordada a web semântica e sua relação com a indexação automática e a recuperação da informação.

A conclusão consiste na realização de uma análise da revisão de literatura oferecida, com uma cronologia dos autores na área de indexação automática. Apresenta-se uma comparação com a tese de doutorado elaborada por Ana Paula Ladeira em 2010, e a partir dessa análise, gráficos e tabelas apresentam os resultados finais.

2 CONSTRUINDO O OBJETO DE ESTUDO E O REFERENCIAL TEÓRICO

2.1 JUSTIFICATIVA

Estudos e experimentos na área de indexação automática surgiram pela necessidade de organização da informação e como evolução nas aplicações de tecnologia da informação em PLN. Desse modo, foram levantados vários trabalhos sobre indexação automática nos últimos anos devido a sua importância para o tratamento, busca e recuperação da informação.

Dois novos conceitos merecem destaque: a internet e a web 3.0, pois apresentam novas perspectivas na recuperação da informação devido ao grande número de conhecimento acumulado, muitas vezes, de forma desorganizada e com difícil acesso.

Com isso, surgiram softwares de indexação automática na tentativa de organizar essa informação de forma mais acessível e prática aos seus usuários. O presente trabalho visa estudar a evolução da indexação automática e seus experimentos ao longo dos últimos 32 anos, de 1981 até o ano de 2013.

2.2 OBJETIVOS DA PESQUISA

2.2.1 Objetivo geral

Apresentar um panorama da literatura científica sobre indexação automática a partir do levantamento de estudos realizados nos últimos 32 anos, em bases de dados nacionais e estrangeiras. Ressalta-se que este trabalho não tem a pretensão de ser exaustivo.

2.2.2 Objetivos específicos

- Verificar na área de Biblioteconomia e Ciência da Informação, os estudos relacionados com a indexação automática;
- Apresentar os experimentos realizados na indexação automática;

- Comparar os resultados obtidos com os apresentados pela pesquisadora Ana Paula Ladeira em 2010, em sua tese de doutorado.
- Apresentar brevemente a indexação automática na recuperação da informação e sua relação com a internet e web 3.0.

2.3 DELIMITAÇÃO DO ESTUDO

Nesse estudo não serão discutidos textos sobre aplicações de indexação automática na área de tradução automática, e questões sobre Linguística Computacional. Sendo estes, os limites temáticos da pesquisa.

O estudo abarcou período de 1981 a 2013, ou seja, os últimos 32 anos de pesquisas e trabalhos acerca da indexação automática. O levantamento foi realizado nas seguintes bases de dados:

- Nacionais: Capes, UNB (catálogo da Universidade de Brasília), USP (Dedalus), Rede Pergamun, Brapci, Unicamp, UFMG (diretamente na base específica da área de ciência da informação), BDTD (Biblioteca Digital Brasileira de Teses e Dissertações);
- Internationais: LISA (Library Information Science Abstract), ARIST (Annual Review of Information Science and Technology), Scielo, Universidade Complutense de Madrid, sites de busca como o Google, entre outras fontes de informação, como livros e obras referenciadas nos materiais.

Importante ressaltar que o trabalho não pretende ser exaustivo, embora tenha buscado englobar o máximo dos resultados encontrados nas pesquisas das referidas bases de dados.

2.4 METODOLOGIA

É importante apresentar o conceito de metodologia para uma melhor compreensão do estudo. Desse modo, a metodologia é uma disciplina que se relaciona com a epistemologia, ou seja, está ligada com a origem e a validade do conhecimento. A metodologia consiste em “estudar e avaliar os vários métodos disponíveis, identificando suas limitações ou não em nível das implicações de suas utilizações” (BARROS, 2000).

De modo geral, a metodologia é utilizada com o intuito de avaliar e examinar as técnicas de pesquisa, além da geração e verificação de novos métodos que conduzem à captação e ao processamento de informações para que se chegue à resolução de problemas de investigação. A metodologia surge então, como um conjunto de procedimentos utilizados na obtenção de determinado conhecimento (BARROS, 2000).

O trabalho consistiu em uma revisão de literatura, propõe-se a apresentar a situação da área de indexação automática atualmente, mostrando o que já foi escrito, e quem escreveu sobre o tema. A revisão de literatura, segundo Santos (2006), tem papel fundamental no trabalho acadêmico , pois por meio dela é possível situar a área de estudo em que o trabalho se encontra, contextualizando-o.

A partir dela, pode-se observar o que já foi pesquisado por outros autores, e o que ainda precisa ser. (ECHER, 2001). Santos (2006) afirma que: “através da revisão de literatura, você reporta e avalia o conhecimento produzido em pesquisas prévias, destacando conceitos, procedimentos, resultados, discussões e conclusões relevantes para seu trabalho”.

Baseado nas afirmações apresentadas acima utilizou-se a tese de doutorado de Ana Paula Ladeira (2010) para a realização de uma comparação entre os resultados obtidos neste trabalho e os resultados encontrados pela autora. Ladeira (2010) realizou um estudo analisando o conhecimento de Processamento da Linguagem Natural encontrado na base ARIST dos anos de 1973 até 2009. A autora utilizou uma amostra de 68 documentos para analisar seu conteúdo apresentando as temáticas mais discutidas pela comunidade científica.

Para a recuperação nos catálogos e nas bases de dados citadas anteriormente, foram utilizados os seguintes termos de pesquisa: “indexação automática”, “processamento da linguagem natural”, “indexação automatizada”, e também os mesmos termos em inglês “automatic indexing”, “natural language processing”. Todavia, outros termos também foram considerados nos resultados, como por exemplo: “indexação semi-automática”.

Foram levadas em conta, as expressões mencionadas que se encontravam no título, resumo, e palavra-chave dos documentos. As buscas abrangem os anos de 1981 até o ano de 2013.

3. Indexação automática: uma revisão de literatura

3.1 Histórico

A indexação surge, segundo Silva e Fujita (2004) com a atividade de elaboração de índices. As autoras afirmam que:

> "(...) a atividade de indexação, como processo, é realizada mais intensamente desde o aumento das publicações periódicas e da literatura técnico-científica, surgindo a necessidade de criação de mecanismos de controle bibliográfico em centros de documentação especializados."

Para Kobashi (1994) a documentação do modo que é hoje, nasceu no século XVII com a edição de *Le Journal des Sçavans* publicado em Paris no ano de 1665. Era um periódico semanal que trazia resumos dos trabalhos científicos, filosóficos e artísticos.

Também de acordo com Collinson (1971, *apud* SILVA; FUJITA, 2004), o primeiro tipo de indexação existente era baseado na memória. E a partir daí passou por muitos séculos de evolução, inclusive na biblioteca de Alexandria. A partir do século XIV era comum a elaboração de catálogos dos livros existentes nos mosteiros. E depois disso surgiram os guias para cada livro.

A indexação surgiu, então, em grande escala em 1737 com a compilação da primeira concordância completa da Bíblia por Alexandre Cruden. (COLLINSON ,1971, *apud* SILVA; FUJITA, 2004).

A literatura mostra que a indexação passou a ter maior atenção com o surgimento dos periódicos, organizando por assunto esse tipo de documento. Contudo, o século XIX, de acordo com Silva e Fujita (2004), foi o período em que a indexação começou a apresentar um aprimoramento de sua execução e a ser apreciada pelo público que via o grande aumento da massa documental.

Foi também no século XIX que os índices evoluíram de forma significativa, partindo de índices de obras isoladas para os índices de vários volumes e para os índices cooperativos e em nível internacional.

Segundo Borges (2009) em sua Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais:

> “Considerando a evolução do tratamento da informação, a indexação pode ser entendida como uma operação de tratamento temático, que comporta as atividades de análise, síntese e representação do conteúdo do documento.”

Desse modo, a indexação nasce como uma das principais áreas da Biblioteconomia e Documentação, e vem da necessidade de se organizar a informação de modo a recuperá-la mais fácil e rapidamente a partir da representação do conteúdo de cada unidade documentária. De acordo com Bastos (1988), a indexação está entre as diversas formas de análise do conteúdo oferecendo uma condensação do assunto do documento.

3.2. CONCEITUAÇÃO

A indexação é definida, de acordo com Cintra (1983 *apud* HOLANDA, 2012), como a tradução de um documento em termos documentários, ou seja, em descritores, cabeçalhos de assunto, os quais têm por função final expressar o conteúdo do documento. É um dos processos básicos de recuperação da informação, podendo ser realizada pelo homem (indexação manual), e também por programas de computador (indexação automática)

Entre as diversas formas de análise de conteúdo, a indexação é a técnica que oferece uma melhor condensação do assunto do documento, atribuindo descritores e aumentando o desempenho na sua capacidade de recuperação por parte do usuário.

Desse modo, a informação é representada por um conjunto de conceitos ou combinações de conceitos selecionados do próprio texto ou de algum vocabulário controlado. Importante ressaltar que a indexação pode fazer uso de termos da linguagem natural, ou como dito anteriormente, de termos convertidos para o vocabulário do sistema.

A indexação se trata de um processo intelectual “ que pressupõe que o acesso à informação documentária, por intermédio dos termos de indexação, será o ponto de partida para selecionar os próprios documentos.” (ROBREDO, 2005). O processo seleciona vários descritores por meio de um tesauro ou lista de cabeçalhos.

Na visão de Robredo (2005), o índice é considerado o instrumento mais importante para a recuperação da informação. É tido como a ponte entre o conteúdo de um documento e os usuários. A indexação consistiria, então, em indicar o conteúdo temático de uma unidade de informação, através da atribuição de termos (um ou mais) ao documento, caracterizando-o.

Holanda (2012) afirma que o objetivo principal da indexação é assegurar a recuperação de qualquer documento ou informação dentro de um sistema de informações. Segundo a autora, a indexação é definida como a tradução de um documento em termos documentários, ou seja, descritores, cabeçalhos de assunto, palavras-chave, com o intuito de expressar o conteúdo do documento.

De acordo com Bastos (1988) a indexação é uma das operações significativas que compõem o ciclo documentário. Indexar é o ato de definir termos ou selecioná-los de modo a descrever o conteúdo do documento da melhor maneira possível para a recuperação da informação. Existem pelo menos duas formas de se analisar o conteúdo de um documento: indexação automática e indexação manual.

Conforme afirma Câmara Júnior (2007), a indexação parte da ideia que a seleção do documento tem como ponto de partida o acesso a informação documentária. Desse modo, a finalidade principal da indexação é a recuperação da informação, satisfazendo os usuários potenciais (ROBREDO, 2005).

A indexação é uma técnica de análise de conteúdo condensando a informação significativa de um documento, por meio da atribuição de termos, gerando assim uma linguagem intermediária entre o usuário e o documento. É tido como um dos processos básicos de recuperação da informação.

De modo geral, todas as definições convergem para o fato de que a indexação tem a função de representar o conteúdo da melhor maneira possível, seja por meio de termos livres ou vocabulários controlados, a fim de torná-lo não só acessível, mas também de fácil recuperação para o usuário que necessita daquela informação.

O principal objetivo da indexação é assegurar a melhor recuperação de qualquer documento ou informação no momento em que houver a solicitação de um usuário em um sistema de informações (HOLANDA, 2012). Segundo a autora, a indexação realiza o registro dos conceitos contidos num documento,

da forma mais organizada e de fácil acesso, por meio da constituição de instrumentos de pesquisa documental como catálogos alfabéticos e índices.

Por meio da indexação é possível se obter um melhor aproveitamento no processo de busca e recuperação da informação, isso se deve ao fato de que o elemento fundamental estabelecido é a representação do conteúdo dos documentos (CÂMARA JÚNIOR, 2007). Indexar é substituir o texto de um documento por uma descrição do conteúdo, tornando possível a recuperação das informações contidas nessa unidade documental.

Robredo (2005), em seu livro *A documentação de hoje e de amanhã*, apresenta as etapas no processo indexação, que seriam as seguintes:

- análise conceitual do conteúdo do documento, ou seja, identificação do assunto; (etapa subjetiva)
- expressão dessa análise, por meio de um conjunto de frases ou palavras; (etapa subjetiva)
- tradução da descrição dos assuntos relevantes para a linguagem de indexação;
- organização das descrições dos assuntos de acordo com a sintaxe da linguagem de indexação.

A etapa referente à identificação do assunto, segundo o mesmo autor, pode ser dividida em outras três etapas, simplificadas a seguir:

- compreensão do conteúdo do documento;
- identificação dos conceitos que representam o documento;
- seleção dos conceitos que poderão servir na recuperação.

De forma sintética, Borges (2008) afirma que indexar é substituir o texto de um documento por uma descrição de seu conteúdo de forma abreviada, com a intenção de apresentar a sua essência.

Cabe ressaltar que dentro da área de indexação (tanto manual quanto automática), é de suma importância se considerar a eficiência da recuperação da informação; desse modo, levam-se em consideração os índices de precisão e revocação do sistema. Lancaster (2004), define revocação como sendo a recuperação de documentos úteis; e emprega precisão como a capacidade de evitar documentos inúteis.

Considerando-se um grande volume de informações, é preferível que se tenha um alto índice de precisão e não de revocação. De modo que nos grandes sistemas de busca, além da recuperação da informação, deve-se ter como objetivo a precisão nos resultados da busca, pois uma grande revocação, geraria muitos resultados a serem examinados pelo usuário (SANTOS, 2009).

Para melhor entendimento e apresentação do tema central desse estudo, serão apresentados a seguir, as definições, diferenças e vantagens da indexação automática e também da indexação manual.

3.3. INDEXAÇÃO MANUAL

A indexação manual é uma tarefa que requer conhecimento do assunto do documento, consistência técnica e desenvolvimento de linguagens de indexação apropriadas a cada sistema de informação (BASTOS, 1988). Essa técnica exige mais tempo do profissional da informação em cada documento analisado, tornando a tarefa muito demorada.

Holanda (2012) afirma que a indexação manual seria a seleção cuidadosa da terminologia empregada, realizada por um indexador que escolhe um conjunto de termos ou combinações para representar o conteúdo do documento.

De acordo com Bastos (1984), a indexação manual requer uma análise intelectual, que compreende basicamente três fases:

- compreensão do conteúdo do documento por meio da leitura completa do texto, título, resumo, entre outras partes que compõem o documento;
- identificação de conceitos, de modo a estabelecer o ambiente lógico;
- seleção dos conceitos, observando a exaustividade, especificidade e consistência.

Segundo Borges (2008), o processo de indexação manual pode ainda ser dividido em duas etapas: a análise conceitual e a tradução. A análise conceitual é relatar o assunto tratado no documento; essa etapa exige a leitura e compreensão do conteúdo, contudo, por demandar muito tempo do indexador, é uma tarefa preocupante.

Ainda diante da análise conceitual, Borges (2008) afirma que é preciso considerar o domínio no qual o documento está inserido, identificando características específicas do campo do conhecimento, de modo que o conhecimento do indexador sobre este domínio (assunto) assume grande importância. Sendo assim, o documento será considerado como um todo, e não apenas como uma parte isolada.

Vale ressaltar que o documento pode ser indexado com exaustividade, ou seja, o indexador realiza a indexação em profundidade, indo além do assunto principal, indexando também assuntos secundários. Também pode-se

optar pela indexação com especificidade, escolhendo precisamente os termos que serão utilizados. Nas definições de Robredo (2005), a exaustividade é o processo que se refere ao nível de reconhecimento dos conceitos ou noções do documento; já a especificidade diz respeito ao nível de abrangência dos conceitos escolhidos.

Foskett (*apud*, BASTOS, 2012) afirma que a exaustividade é a extensão com que se realiza a indexação de um documento, com o intuito de estabelecer todos os assuntos que esse documento trata; já a especificidade é a extensão em que um sistema de informação permite ser preciso ao especificar o assunto de determinado documento.

A etapa referente à tradução consiste em converter o que foi analisado no documento em termos de indexação, ou seja, o indexador deve selecionar os termos que representam adequadamente o assunto do documento. Essa etapa pode ser realizada com o auxílio de um vocabulário controlado. Entre eles estão: taxonomia, tesauro, lista de cabeçalhos de assunto.

Segundo Lancaster (2003, p. 286): “os indexadores humanos procurarão selecionar expressões do texto que pareçam ser bons indicadores daquilo de que trata um documento.“

3.4 PROBLEMÁTICAS DA INDEXAÇÃO MANUAL

Por se tratar de uma tarefa intelectual é natural que existam problemas e divergências entre os indexadores. Desse modo um mesmo documento pode ser indexado de formas diferentes por pessoas diferentes (inconsistência interindexadores); ou até pela mesma pessoa que se encontra em momentos distintos (inconsistência intraindexador).

Teoricamente, o indexador humano deveria produzir uma indexação dita superior em relação à realizada por programas de computador, contudo, na prática, a indexação manual apresenta muitos problemas e inconsistências, afetando, diretamente, a recuperação da informação.

Lancaster (2004) afirma que uma mesma publicação pode apresentar conjuntos diferentes de termos de indexação, dependendo do grupo de usuários ao qual se destina e dos interesses particulares desse grupo. Isso não

quer dizer que a indexação de um ou de outro estará errada, são apenas maneiras distintas de indexar.

Outra questão apresentada por Borges (2009) diz respeito ao fato de que durante a indexação manual, o indexador responsável pela leitura documental realiza grandes esforços cognitivos, como identificação dos pontos mais importantes, pausas para reflexão do texto entre muitos outros. Isso acaba sendo a parte mais cansativa da indexação.

Cabe ressaltar que o indexador, de forma geral, não dispõe de muito tempo, e por isso não pode se dedicar por horas a leitura de um só documento. Outro fator é que a indexação requer tempo e exige conhecimentos adequados do indexador (ROBREDO, 2005); tornando-a uma operação cara. Desse modo, a indexação automática entra como um grande auxílio, realizando uma extração inicial de termos.

Um terceiro aspecto de suma importância apresentado na indexação manual, diz respeito à subjetividade, ou seja, o envolvimento humano no ato de indexar, de modo que o nível de conhecimento do indexador influencia a atividade e a consistência da indexação do documento, fazendo com que a experiência do indexador acabe por interferir no processo.

A indexação manual, de acordo com Borges (2008), ainda consiste em um procedimento caro, uma vez que é necessário gasto maior com pessoal para a realização da atividade.

3.5. INDEXAÇÃO AUTOMÁTICA

3.5.1 Histórico

A indexação automática surgiu pela necessidade de se criar um sistema capaz de “substituir um especialista humano, mantendo total relação com o conceito de inteligência artificial” (BASTOS, 1984). Com a explosão informacional, o número de documentos cresceu exponencialmente, fazendo com que a indexação manual de todo o material bibliográfico disponível se tornasse inviável.

O principal objetivo da aplicação da indexação automática é que ele possibilite ao usuário ter acesso aos documentos de que necessita, sem a interferência direta de um documentalista.

Segundo Guedes (1994), a indexação automática é a mecanização do processo de indexação, seja em parte ou no todo, tendo o objetivo principal de reduzir de forma significativa a subjetividade encontrada no processo feito manualmente.

O início da indexação automática remete à 1948, quando George Zipf formulou duas leis sobre distribuição de palavras em um texto. A primeira se referia às palavras de alta frequência, na qual Zipf afirmava que ao se colocar as palavras de um texto longo em ordem decrescente de frequência, seria possível verificar que a ordem de série das palavras (R) multiplicada por sua frequência (F) produziria uma constante K, portanto: R x F = K (MAMFRIM, 1991).

Já a sua segunda lei, se referiu a palavras de baixa frequência. Esta lei foi aperfeiçoada por Booth, e ficou conhecida como lei de Zipf-Booth, demonstrada da seguinte forma:

$$\frac{I\,n}{I\,1} = \frac{2}{n\,(n+1)}$$

Onde:

- In é o numero de palavras que ocorreram N vezes para n < 5 ou n < 6;
- I 1 é o número de palavras que ocorreram uma única vez
- 2 é uma constante atribuída a língua inglesa.

As leis apresentadas acima, foram constatadas empiricamente, ou seja, por meio de testes, contudo, não se aplicam em sistemas de informação. Apesar disso, através de uma observação de que as duas leis operam apenas em relação aos extremos da distribuição de palavras em um texto, Goffman, citado por Mamfrim (1991) sugere a existência de um ponto onde haveria a transição de palavras de alta frequência para as palavras de baixa frequência,

ou seja, o número de palavras vai tender para a unidade. Neste ponto, se encontrariam as palavras representativas do conteúdo do documento em questão. (MAMFRIM, 1991)

A partir disso, começou a existir a possibilidade de aplicação das chamadas leis bibliométricas, as quais trabalham com a frequência das palavras para a indexação em sistemas de informação. Este ponto é denominado ponto T, representado da seguinte forma:

$$T = \frac{-1 + \sqrt{1 + 8.I1}}{2}$$

Onde:

- I1 é o número de palavras que ocorreram uma única vez;
- 8 é uma constante derivada da língua inglesa;
- 2 é uma constante matemática da fórmula de Baskara, para resolução de equações de 2º grau.

Segundo proposto por Goffman e apresentado por Mamfrim (1991), uma vez que o ponto T fosse identificado, seria definida uma região dentro da qual estariam as palavras que identificariam o conteúdo do documento, levando em conta a palavras de maior conteúdo semântico (GUEDES, 1994). De acordo com a explicação de Mamfrim (1991), ocorreria o seguinte:

> "(...) Esta região seria definida a partir de um ponto correspondente a uma frequência aproximada. Assim, a partir desta frequência são contadas as palavras entre o ponto T e a palavra de maior frequência. Este mesmo número de palavras é projetado para abaixo do ponto T, definindo uma região."

No Brasil houve algumas tentativas com de aplicação do ponto T à indexação, de modo que com base nos resultados obtidos é possível afirmar que a Fórmula de Transição de Goffman pode ser aplicada à língua portuguesa sem maiores problemas, comprovando que o algoritmo proposto se adequou à língua.

Segundo Guedes (1994), Hans Peter Luhn, por volta da década de 50, foi o precursor da área de estudos bibliométricos fundamentados em frequência de ocorrência de palavras. Luhn propôs que a frequência das palavras em um documento ou conjunto de documentos estaria relacionada com sua pertinência ao processo de indexação, fornecendo medida útil de sua importância. (MAMFRIM, 1991).

Luhn volta seus estudos para classificação e busca automática dos documentos fazendo uso de uma abordagem estatística, além de um método automático probabilístico, visando à criação de resumos. (GUEDES, 1994).

Como citado por Narukawa, Leiva e Fujita (2009), para Luhn a frequência das palavras em um texto tem relação direta com a utilidade destas palavras na indexação, expressando quais são as palavras representativas do conteúdo do documento. Em contrapartida, Baxandale (*apud* GUEDES, 1994) surge analisando "comparativamente a eficiência de três métodos automáticos de indexação de artigos técnicos".

Por meio dessas aplicações foi possível concluir que da quantidade de palavras surge um conjunto de palavras de "qualidade", ou seja, palavras de conteúdo semântico significativo e representativo para o artigo.

3.5.2 Conceituações de indexação automática

Robredo (1986), considera que a indexação automática é qualquer procedimento que permita identificar e selecionar os termos que representem o conteúdo dos documentos, sem a intervenção direta do documentalista/ indexador.

Bastos (1988) define indexação automática como uma técnica que prescinde de certa forma, da presença do homem para a realização intelectual da atividade. Ela visa agilizar e auxiliar o processo intelectual realizado pelos profissionais da área.

Segundo a mesma autora, o processo de indexação automática se refere a uma operação que identifica, por meio de programas de computador, palavras ou expressões que sejam significativas dos documentos, para então, descrever o seu conteúdo de forma condensada.

Robredo (1982) afirma também que o processo de indexação automática se desenvolve seguindo um esquema semelhante ao processo de leitura-memorização, onde há uma memorização temporária que conserva as palavras significativas, modificando ou aperfeiçoando o conceito das mesmas a partir da percepção de novos conceitos significativos.

E há também uma memória permanente dos conceitos, denominada simplificadamente de memória. No fim do procedimento, encontram-se na memória uma série de "palavras-conceitos-descritores" que representam as ideias básicas do documento em questão (ROBREDO, 1982).

Na visão de Holanda (2012, p. 42-59) no processo de indexação automática cada palavra presente no texto tem a capacidade de estabelecer uma entrada no índice, sendo desempenhada por um sistema computacional.

Os descritores ou palavras-chaves, são escolhidos do título, do resumo, ou até mesmo do próprio texto do documento, seguindo uma série de regras, as quais verificam sua validade como descritores, por meio da comparação com os termos de tesauros ou dicionários adequados. (ROBREDO, 2005)

Neves (2009) reafirma que a indexação automática seria a execução de um processo de representação de documentos, porém, realizada por meio de programas ou algoritmos de computador que "varrem" o documento e realizam a representação do conteúdo sem a intervenção direta do indexador.

Ainda que o termo mais utilizado seja 'indexação automática', algumas técnicas não são de fato totalmente automatizadas, são chamas de semi-automáticas, como a *Machine-Aided indexing*, que utiliza um vocabulário controlado, e realiza a comparação entre as expressões extraídas de um documento e as de uma linguagem documentária. Por outro lado, exemplos de técnicas de processamento totalmente automático são a " Categorização de texto" e o "Agrupamento" (*clustering*). (NEVES, 2009 apud HJORLAND, 2008).

Conforme Golub (2005, *apud* NEVES, 2009) existe uma diferença entre as técnicas citadas acima. A categorização de texto envolve a construção de indexadores automáticos capazes de aprender e classificar documentos tendo como apoio um conjunto de categorias pré-definidas e documentos pré-classificados manualmente, os quais servem de modelo para o sistema aprender e classificar um novo documento.

Maia e Souza (2010) afirmam que a técnica do agrupamento, segundo Maia e Souza (2010) permite subdividir um conjunto de objetos em grupos e não faz uso de categorias pré-definidas ou documentos pré-classificados; suas relações derivam automaticamente de documentos a serem agrupados e inseridos nos *'clusters'* respectivos.

O objetivo é fazer com que cada *cluster*, ou seja, grupo, se torne o mais homogêneo possível, levando em consideração as similaridades dos objetos dos grupos, e que os objetos sejam diferentes. (MAIA; SOUZA, 2010).

Segundo Borges (2008) a indexação automática pode ainda ser dividida em indexação por extração automática e indexação por atribuição automática.

3.5.3 Indexação por extração automática

No processo de indexação por extração automática palavras ou expressões que aparecem no texto são extraídas para representar seu conteúdo como um todo. Caso se trate de uma versão eletrônica do documento é possível ainda utilizar um software para extrair os termos a partir de princípios utilizados também por seres humanos, como a frequência e a posição da palavra, e o próprio contexto onde ela se insere.

Borges (2008) afirma que os sistemas baseados em indexação por extração automática realizam as seguintes tarefas:

- contar palavras num texto;
- cotejá-las com uma lista de palavras proibidas;
- eliminar palavras não-significativas (artigos, preposições, conjunções, etc.);
- ordenar as palavras de acordo com sua frequência.

Desse modo, a indexação por extração automática é aquela realizado por meio dos termos encontrados no próprio texto, onde são extraídas as palavras consideradas mais representativas do documento.

3.5.4 Indexação por atribuição automática

O processo de indexação por atribuição automática é mais complexo de ser realizado, pois está atrelado ao controle terminológico. Segundo Lancaster

(2004), esse tipo de indexação “envolve a representação do conteúdo temático por meio de termos selecionados de algum tipo de vocabulário controlado”.

Em complemento Borges (2008) pontua que para cada termo atribuído, conjuga-se um determinado ‘perfil’ de palavras ou expressões que por ventura ocorram nos documentos. Isto salienta uma relação semântica entre as palavras, atribuindo outros conceitos aos termos escolhidos, como no exemplo: ‘chuva ácida’, pode-se atrelar as expressões ‘ poluição atmosférica’ e ‘ precipitação ácida’. Desse modo é indexado o conteúdo do documento e também termos que possam estar relacionados com o assunto tratado, fazendo com que na hora da busca, o usuário tenha outros termos para utilizar.

3.5.5 Programas de geração de índices na indexação automática

A indexação automática pode estar relacionada com o uso de programas computacionais para a geração de índices pré-coordenados. Segundo Lancaster (2004, p. 52), “vários programas de computador foram desenvolvidos para gerar automaticamente um conjunto de entradas de índice a partir de uma sequência de termos.” Como exemplo podem-se citar o KWIC, o KWOC e o KWAC.

Todos são métodos simples para a construção de índices a partir de texto. O KWIC (*Key-word-in-context* ou Palavra-chave no contexto), segundo Borges (2008):

> “foi desenvolvido em 1959 e diz respeito a um índice rotativo em que cada palavra-chave que aparece nos títulos dos documentos torna-se uma entrada do índice. Cada palavra-chave é destacada de alguma forma e as palavras restantes do título aparecem envolvendo-a”.

O critério usado para selecionar as palavras é o seguinte: o programa reconhece as palavras que não são palavras-chaves, baseando-se em uma lista de palavras proibidas ou vazias (artigos, preposições, conjunções), e então, ele compara as palavras do título com a referida lista, ignorando aquelas que constarem na mesma, ou incorporando as que não o são.

Se trata de um método que não faz uso de tesauros ou dicionário, de modo que a lista de termos “significativos” não tem nenhum controle, realizando a indexação por meio de linguagem natural. Isso acaba por gerar alguns problemas, como a não-identificação dos sinônimos. Além disso, alguns

termos tidos como significativos podem, muitas vezes, ter pouco valor, aumentado o “nível de ruído do índice.” (ROBREDO, 2005).

Apesar de ser um método barato e de fácil utilização, ele está diretamente relacionado à qualidade dos títulos, considerando que estes sejam bons indicadores do conteúdo dos textos. (LANCASTER, 2004, p. 54.-55).

A respeito do método KWOC (*Key-word-ou-of-context* ou Palavra-chave fora do contexto), ele é bastante semelhante ao KWIC, contudo as palavras-chave que se tornam pontos de acesso são repetidas fora do contexto, normalmente destacadas no canto esquerdo da página ou usadas como cabeçalhos de assunto (BORGES, 2008, p. 185).

Vale ressaltar que no KWOC as palavras extraídas são separadas das outras palavras que consituem parte do documento, sendo substituídas por algum sinal gráfico, tornando difícil a recuperação de termos compostos. (NARUKAWA, 2011)

Existe ainda o índice KWAC, chamado ‘ keyword and context’ ou seja, palavra-chave e contexto, esse índice não é muito diferente do índice KWOC apresentado anteriormente. Porém, segundo Narukawa (2011), enquanto no índice KWOC, o lugar que a palavra extraída ocupava no título é indicado por um sinal gráfico, no KWAC, a palavra extraída permanece na parte considerada.

Nas décadas de 60 e 70, surgiram outros sistemas relacionados à indexação automática, destacando-se o SMART e o MEDlars. O SMART funciona sem análise manual do conteúdo. Trechos do documento são introduzidos no computador e uma variedade de procedimentos automáticos de análise de texto é utilizada para produzir para cada item um ‘conceito vetor’ constituído por ponderação de termos ou conceitos representativos do conteúdo do documento (SALTON, 1968 *apud* BORGES, 2009).

O MEDlars por sua vez, faz uso de vocabulário controlado, onde a recuperação é efetuada por meio de uma comparação entre uma lista de palavras-chave determinada para os documentos com os termos de busca das formulações booleanas de pesquisa (SALTON, 1968 *apud* BORGES, 2009).

Contudo, apesar dos esforços e das grandes vantagens que a indexação automática pode oferecer aos indexadores e aos usuários, ela enfrenta óbices ao não reconhecer o processo mental de análise de assunto envolvido no

processo de indexação. Desse modo ela não representa os assuntos do documento com a mesma efetividade atribuída à humana.

Segundo Ward (1996, *apud* SILVA e FUJITA, 2004) a indexação automática é incapaz de fazer relações entre textos ou entre textos e uma visão de mundo; é limitada ao vocabulário controlado e não consegue indexar o que está implícito.

3.5.6 Análise probabilística

A indexação automática pode partir de análises estatísticas (ou probabilísticas) das ocorrências das unidades léxicas, com o intuito de provar que a frequência das palavras pode expor o que é realmente importante no texto.

Desse modo, a inclusão de estruturas e cálculos matemáticos e estatísticos são usados para atribuir peso às palavras do texto (TAVARES JUNIOR, 2006). Assim, torna-se possível criar um mecanismo mensurável para escolher descritores a partir das palavras mais relevantes para se representar o assunto do documento.

Segundo Lancaster (2004), ao invés da frequência absoluta, deve-se utilizar a frequência relativa para selecionar os termos. Por meio desse método, deverão ser selecionadas as palavras ou expressões que ocorram num documento com mais frequência do que sua ocorrência na base de dados como um todo. Isso se torna um pouco complicado porque exige que se faça uma contagem da frequência pela qual cada palavra ocorre na base de dados e também uma comparação dessa ocorrência com a de uma palavra em determinado documento.

Lancaster (2004) ainda ressalta que os termos obtidos da frequência relativa não serão totalmente diferentes daqueles da frequência absoluta, uma vez que os termos novos serão os que ocorrem raramente no documento e na base de dados. Os termos que desaparecerão serão aqueles que ocorrerem frequentemente na base como um todo.

3.5.7 Análise linguística

Segundo Gil Leiva (1999, *apud* NARUKAWA, LEIVA, FUJITA, 2009), a partir do início da década de 1960, surgem estudos associando as técnicas de processamento de linguagem natural à automatização da indexação; seguindo em direção a compreensão da estrutura textual, suas relações e significados.

Na perspectiva de Medeiros (1999, *apud* CÂMARA JÚNIOR, 2007), os componentes utilizados no processamento de linguagem natural executam tarefas de reconhecimento do texto segundo o nível de conhecimento linguístico exigido ao tratamento. Atuando em níveis de profundidade diferentes, e também um grau de dificuldade de implementação crescente .

São quatro componentes apresentados a seguir (CÂMARA JUNIOR, 2007):

- componente morfológico: se preocupa com a maneira como as unidades léxicas são apresentadas. Faz uso de um dicionário, para identificar as palavras válidas na linguagem utilizada.
- componente sintático: é responsável pela organização das orações. Por meio da sintaxe é possível reconhecer a estrutura das frases e as funções de seus componentes.
- componente semântico: visa analisar as frases sintaticamente corretas para avaliar se são compreensíveis, formalizando a interpretação do texto.
- componente pragmático: procura incluir o contexto a análise linguística, a fim de permitir a geração de um significado.

3.5.8 Redes semânticas e neurais

As redes semânticas são estruturas que representam relações entre conceitos. Essas relações são denominadas axiomas ou asserções, e permitem inferir conclusões a partir da rede semântica. Desse modo, “as redes semânticas são úteis para a representação do conhecimento em vários domínios”. (LUCENA, 2003)

Conforme apresentado por Salinas Ordoñez e Gelbuk (2010), as redes semânticas surgem a partir de trabalhos linguísticos apresentados em 1968, **e** a partir de diferentes contribuições se consolidaram na década seguinte.

Segundo os mesmos autores, existem duas tendências: de um lado as redes estruturadas e os sistemas de representação do conhecimento, e do outro, as multiredes voltadas para as ciências cognitivas.

Alonso Fernández (1993) afirma que as principais características das redes semânticas são:

- sua estrutura de representação permite a organização hierárquica dos dados com a possibilidade de busca seletiva destes dados, para serem utilizados pelos mecanismos de busca.
- sua localização ou situação topográfica dos conceitos na rede é significativa, assim como as relações de proximidade entre os conceitos.
- alguns sistemas permitem ter em mente o contexto e os diferentes pontos de vista acerca deste mesmo contexto
- a uniformidade
- a capacidade de manipulação de suas estruturas de conhecimento, é do ponto de vista computacional, umas das propriedades mais importantes deste sistema.

De acordo com Brachman (*apud* ALONSO FERNANDÉZ, 1993), a semântica da rede consiste em sua capacidade de representar a semântica das expressões em linguagem natural. Desse modo, a rede semântica constitui uma ordem onde seu valor reside na inter-relação de seus constituintes e depende da coerência de sua formulação.

Em relação às redes neurais, Alonso Fernández (1993) acrescenta que estas redes são inspiradas em modelos biológicos do funcionamento do sistema nervoso do ser humano e suas abordagens são rigorosamente matemáticas. Os estudiosos tentam construir um modelo de computador que imite os processos biológicos do cérebro humano (ALONSO FERNÁNDEZ, 1993).

Ainda de acordo com o autor, um sistema computacional de redes neurais tenta construir um grande número de unidades de processamento básicas para configurar o nível pré-simbólico dos processos cognitivos que se pretende simular no computador. As unidades de processamento se constroem

de forma semelhante à estrutura e funcionamento de um neurônio do sistema nervoso humano.

Ademais, Silva e Fujita (2004) pontuam que o indexador faz uso de aspectos cognitivos que interagem na leitura. Desse modo, os processos cognitivos utilizados pelo leitor são os seguintes: o seu conhecimento sobre a estrutura textual, o conhecimento prévio sobre o assunto, a recuperação de esquemas formados com sua experiência de vida, fazendo com que o leitor faça inferências sobre o assunto abordado.

Segundo Naves (2000, *apud* SILVA; FUJITA, 2004), o processo de inferência se divide em:

- inferência lógica – estabelecendo causas, motivações;
- inferência integrativa – baseia-se nos conceitos e propriedades da organização hierárquica;
- inferência construtiva – refere-se ao conhecimento do indexador.

Desse modo Shaw e Fonchereaux (1993, *apud* SILVA; FUJITA, 2004) afirmam que cabe ao indexador, no momento da análise documentária, decidir sobre outros dois aspectos cognitivos:

1. decidir sobre o que o texto fala
2. traduzir essa decisão em termos de indexação

Nesse contexto as redes neurais utilizam suas abordagens voltadas para modelos biológicos na tentativa de construir uma máquina que represente os processos do ser humano.

3.6 Linguística computacional

A linguística computacional é a área que se dedica à compreensão da língua e de técnicas apropriadas à sua interpretação, seja nas modalidades escrita ou falada, tentando imitar a capacidade humana de comunicação. (BORGES, 2008).

Desse modo, essa área faz uso dos elementos da sintaxe, semântica, fonética, fonologia, pragmática e análise do discurso. É a parte da linguística

onde os algoritmos são aplicados a coleções de material de linguagem. (LADEIRA, 2010).

Haller (1983) afirma que, na linguística computacional, a análise linguística surge como um instrumento fundamental para aumentar de forma significativa a capacidade de um sistema automatizado de armazenamento e recuperação de informação.

Segundo Gil Leiva e Rodriguez Muñoz (1996), a linguística computacional é a interseção entre a linguística e a informática com a finalidade de processar/gerar línguas. Ou seja, é a área que explora as relações entre linguística e informática.

Diferentemente de um ser humano, para um sistema computacional, um texto escrito em linguagem natural, corresponde a uma cadeia de símbolos sem significado algum. Com isso, para que seja possível a compreensão por parte do computador, é necessário que se recorra a técnicas próprias de PLN juntamente com a linguística computacional (SALINAS ORDOÑEZ e GELBUKH, 2010).

A área da linguística computacional pode ser dividida em linguística de *corpus* e processamento da linguagem natural. A linguística de *corpus* trabalha com o "*corpora* eletrônicos", ou seja, "grandes bancos de dados que contenham amostras de linguagem natural" (BORGES, 2008). Desse modo, o objetivo não é produzir um software, mas, sim, estudar os fenômenos linguísticos que podem acontecer em grandes amostras de uma determinada língua.

O processamento da linguagem natural, por sua vez, visa o estudo da linguagem diretamente voltado para a construção de softwares, como *parsers*, tradutores automáticos, *chatterbots*, reconhecedores automáticos de voz, entre outros.

Um *parser* pode ser assim definido:

> [...] um *parser*, no contexto da linguística computacional é um analisador automático (ou semi-automático) de sentenças [frases]. Esse tipo de programa é capaz de analisar uma sentença com base em uma gramática preestabelecida de determinada língua, verificando se as sentenças fazem parte ou não da língua, de acordo com o que autoriza a sua gramática. Um *parser* também analisa sintaticamente as sentenças [...] (OTHERO; MENUZZI, 2005 *apud* BORGES, 2009)

Os *chatterbots* por sua vez, são programas desenvolvidos para interagir com usuários humanos por meio de diálogo em linguagem natural, na forma escrita. Com isso cabe a área de PLN a construção de programas que sejam capazes de interpretar/gerar informações em linguagem natural. (OTHERO, 2006).

Acerca da utilização da gramática pela linguística computacional, Borges (2008) entende que:

> "Dentro da área da linguística computacional, a gramática possui o significado de um conjunto de regras e vocábulos de uma língua, conjunto este relativamente pequeno, e que possibilita, por sua vez, reconhecer todas as frases possíveis de uma determinada língua, desse modo, se atribui a essas frases uma estrutura sintagmática, construindo uma espécie de 'gramática sintagmática". (BORGES, 2008)

De acordo com Conteratto (2006), um dos fatores que implica diretamente na eficiência de um sistema computacional é o fato do software possuir uma descrição linguística suficientemente informativa e organizada, demonstrando a importância do estudo da semântica para a eficiência dos sistemas de PLN.

3.7. Processamento da linguagem natural

O processo de indexação requer conhecimentos sobre o processamento da linguagem natural. Segundo Neves (2009), é na década de 1960 que se dá o início da aplicação de técnicas na área de processamento da linguagem natural para a indexação automática.

A recuperação por meio de linguagem natural tem sido "o apoio mais concreto para os recentemente criados motores de busca na web" (LADEIRA, 2010). O processamento da linguagem natural pode ser definido como qualquer utilização do computador para a manipulação da linguagem natural

Nesse campo é estudado como o computador pode ser usado para processar dados de linguagem, com o intuito de promover a reorganização, a extração e a construção de sentidos. Contudo, "é preciso considerar os componentes que intervém na linguagem – morfológicos, lexicais, sintáticos, semânticos, lógicos" (BRITO, 1992).

Segundo Conteratto (2006), os sistemas de processamento da linguagem natural são modulares em sua maioria, nos quais os diferentes níveis de processamento (morfológico, sintático, semântico, discursivo e pragmático) são executados em diferentes módulos. Com isso, tem-se o texto como entrada e uma representação formal do mesmo como saída.

Assim, um sistema de processamento de linguagem natural, pode começar no nível da palavra para determinar a estrutura morfológica e, em seguida, passar para o nível da frase, determinando a ordem das palavras e o significado da frase completa, partindo, então, para o ambiente geral.

A palavra ou frase pode ter um significado específico em um determinado contexto, além de poder estar relacionada com outras palavras do restante do texto. Dessa forma, torna-se importante uma visão ampla, classificando o conhecimento por um sistema em alguns níveis:

- nível fonético: lida com a pronúncia;
- nível morfológico: lida com as menores partes da palavra;
- nível lexical: lida com o significado lexical das palavras e partes de análise de discurso;
- nível sintático: lida com a gramática e a estrutura das frases;
- nível semântico: lida com o significado e o sentido das palavras e frases;
- nível de discurso: trata da estrutura de diferentes tipos de textos;
- nível pragmático: lida com o conhecimento que vem do mundo exterior.

Assim, um sistema de processamento de linguagem natural pode envolver alguns ou todos esses níveis de análise.

Cabe ressaltar que relacionado ao conceito de linguagem natural está o conceito de linguagem documentária, principalmente pelo fato de estas últimas serem utilizadas no processo de representação do conhecimento ao descrever seu conteúdo em um processo parametrizado.

De acordo com Lara (2004, *apud*, LIMA; BOCCATO, 2009), a linguagem documentária é um tipo de linguagem artificial, construída com a finalidade de facilitar e organizar o acesso à informação.

Guinchat e Menou (1994, apud LADEIRA, 2010), acrescetam que as linguagens documentárias são usualmente usadas no momento da entrada de dados dos sistemas de informação, entrando na etapa de análise conceitual e tradução.

De acordo com Lara (2004, apud LIMA e BOCCATO, 2009), a linguagem documentária é um tipo de linguagem construída com a finalidade de facilitar e organizar o acesso a informação, assim como sua transferência.

As linguagens documentárias podem ser definidas como linguagens que foram construídas de símbolos que serão utilizados não só para representar o conteúdo do documento, mas também para armazenar e recuperar a informação. Surgem como um meio de evitar ambiguidades. (SANTOS, 2009).

Dentre os principais tipos de linguagens documentárias estão: sistemas de classificação, cabeçalhos de assunto, palavra-chave, lista de descritores, tesauros, e etc.

Ressalta-se ainda que outro tema importante no processamento da linguagem natural, são as ambiguidades, gerenciar de "maneira computacionalmente eficiente e psicologicamente plausível" (LADEIRA, 2010).

3.7.1 Ambiguidades

Segundo Brascher (2002), ambiguidade diz respeito a "uma expressão da língua (palavra ou frase) que possui vários significados distintos, podendo, consequentemente, ser compreendida de diferentes maneiras por um receptor".

É um fator que pode causar ruído na recuperação da informação ao apresentar ao usuário resultados não condizentes com o sentido requerido da expressão de busca proposta.

Acerca da tipologia das ambiguidades, Fuchs (1996, *apud* BRASCHER, 2002) apresenta a seguinte classificação:

- ambiguidade morfológica: quando não é possível determinar a classificação gramatical de determinada forma; Ocasionada por policategorização, onde palavras pertencem a mais de uma categoria gramatical;

- ambiguidade lexical: quando existe a possibilidade de mais de uma interpretação do significado de uma unidade lexical. É provocada por homografia (palavras iguais com significados diferentes. Ex: cobre, espécie de metal ou, flexão do verbo cobrir) e holissemia (uma só expressão possui significados distintos. Ex: arquivo, espécie de móvel ou instituição que administra um conjunto de documentos);
- ambiguidade sintática: ocorre na estruturação da frase em constituintes hierarquizados;
- ambiguidade predicativa: ocorre na interpretação das relações temáticas entre predicado, argumentos e participantes;
- ambiguidade semântica: quanto existe mais de uma interpretação possível para os termos relacionados na frase;
- ambiguidade pragmática: relacionada com o cálculo dos valores enunciativos, e a situação do falante no momento da frase;

Importante mencionar que o nível pragmático permite novas soluções para problemas de ambiguidade na semântica e na sintaxe, mostrando que a comunicação em linguagem natural depende de condições que vão além desta linguagem propriamente dita (BARANOW, 1983).

Solucionar ambiguidades em sistemas de recuperação da informação visam determinar quais escolhas são mais adequadas, considerando o contexto em que a ambiguidade ocorre (BRASCHER, 2002). Esses sistemas aplicam diferentes técnicas de tratamento automático da linguagem natural e usam regras formais segundo a abordagem linguística e o modelo de representação do conhecimento escolhidos pelo sistema.

3.7.2 Aplicações do PLN na documentação

Segundo Gil Leiva e Rodriguez Muñoz (1996), o processamento de linguagem natural possui aplicações gerais e específicas na área da Documentação, as quais destacam:

- a busca nas bases de dados em linguagem natural. Desse modo, as consultas são mais simplificadas, proporcionando a busca por meio de termos totalmente naturais;

- a geração automática de tesauros, possibilitando a identificação de relações sintáticas e semânticas entre palavras e frases;
- difusão de informação, por meio de um programa que conheça as técnicas de PLN;
- elaboração automática de resumos;
- indexação automática de documentos.

Desse modo, o processamento de linguagem natural pode contribuir de forma eficiente para o tratamento e recuperação da informação, abarcando os conceitos de sintaxe e semântica. Além disso, ele surge como um grande fator para os conceitos de indexação automática.

3.8. As relações entre a sintática e a semântica na indexação

Sabe-se que a semântica e sintaxe possuem juntas papeis fundamentais na indexação automática, pois elas permitem ao sistema que este identifique a estrutura lexical das frases e o significado dos termos que estão representando o conteúdo do documento.

3.8.1 Sintaxe

A palavra sintaxe, significa ordem, combinação, relação, sendo oriunda do grego *sýntaxis.* É entendida como a parte da gramática:

> "que se preocupa com os padrões estruturais dos enunciados e com as relações recíprocas dos termos nas frases e das frases no discurso, enfim, com todas as relações que ocorrem entre as unidades linguísticas no eixo sintagmático" (SAUTCHUK, 2010).

A sintaxe possui suas leis, as quais promovem, autorizam ou recusam certas construções, classificando-as em "pertencentes à língua portuguesa" ou "não pertencentes", de modo que as pertencentes formarão frases aceitas, tornando possível a capacidade de comunicação dos textos. A análise sintática consegue determinar de forma clara e concisa se uma expressão ou frase está adequada a gramática dessa língua específica.

A sintaxe dedica-se à correção das construções verbais numa língua, enquanto a semântica, o seu sentido. Dessa forma, podem existir frases sintaticamente corretas, porém sem nenhum conteúdo semântico aceitável e vice-versa.

A análise sintática trata das frases e dos discursos dos sintagmas e não das palavras, de modo que a análise morfológica cuida das palavras isoladamente. Sintagmas seriam então "expressões que ditam uma relação de dependência, na qual um elo de subordinação é estabelecido e cada um dos elementos é também um sintagma" (BORGES, 2009).

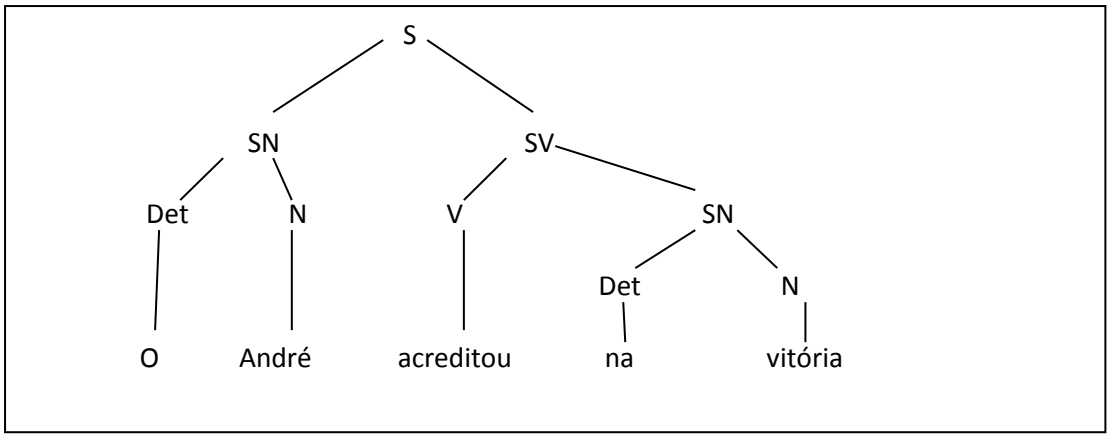
O sintagma pode ser: nominal (nome/substantivo), adjetival (adjetivo), verbal(verbo), preposicional(preposição), e adverbial (advérbio); a sua determinação é muito importante para a análise sintática. É possível, por exemplo, que uma mesma frase possua dois sintagmas do mesmo tipo. Para

determinar um sintagma é importante conhecer e identificar o seu elemento núcleo, que por sua vez, pode ser composto por mais de uma palavra.

Segundo Baranow (1983), a sintaxe das linguagem naturais tem como objetivo “classificar de modo explícito os enunciados contidos em textos falados ou escritos, em corretos ou incorretos, atribuindo-lhes descrições estruturais”.

A seguir um exemplo baseado em Borges (2009), com o intuito de clarificar a ideia de sintagma:

Exemplo: O André acreditou na vitória



Legenda:
S = sentença (frase)
SN = sintagma nominal
Det = determinante
N = nome ou substantivo
SV = sintagma verbal
V = verbo

Fonte: Adaptado de Borges (2009, p. 186).

De modo geral, para que um software de indexação automática funcionar de forma satisfatória, ele deve analisar essencialmente um texto tanto pelo aspecto semântico quanto pelo aspecto sintático. Com o intuito de auxiliar nessa empreitada, surgiram as linguísticas computacionais, vistas anteriormente (Seção 3.6).

3.8.1.1 Sintagmas nominais

De acordo com Kuramoto (1995, *apud* SILVA *et. al*, 2011), o sintagma nominal pode ser definido como a menor parte do discurso portadora da informação. Ainda segundo os autores, em um sintagma nominal, os signos linguísticos "ligam-se uns aos outros formando grupos ao redor de substantivos".

Segundo Miorelli (2001, *apud* SILVA *et. al*., 2011), os sintagmas nominais podem ser entendidos de forma sintática ou semântica. O uso de sintagmas nominais em bases de dados textual para acesso à informação surge como uma alternativa aos sistemas de recuperação da informação.

De modo geral, os sintagmas nominais são retirados do texto e analisados com o intuito de facilitar o procedimento da indexação automática. Vale ressaltar que os sintagmas nominais quando extraídos do texto conseguem manter o seu significado, fazendo com que possam ser utilizados no processo de indexação, o que não ocorre com as palavras,.

Segundo Kuramoto (2002, *apud* SILVA *et. al*, 2011), os sintagmas podem ser utilizados no processo de recuperação da informação de duas formas distintas. A primeira está diretamente ligada à indexação automática, em que no lugar de índices contendo palavras far-se-ia uso de índices contendo sintagmas. A segunda alternativa se refere a uma organização hierárquica em árvore de sintagmas nominais.

A vantagem dos sintagmas sobre as palavras dá-se à medida que as palavras não conseguem eliminar inconvenientes como a sinonímia e a polissemia, fazendo com o que usuário encontre documentos diferentes do que havia sido solicitado.

3.8.1.2 Uso dos sintagmas nominais no processo de recuperação da informação

O processo de recuperação da informação com a utilização de palavras isoladas que fazem parte de um índice para indexação de documentos, não é capaz de suprir as necessidades do usuário, pois não consegue atingir a dimensão semântica dos documentos.

Segundo Kuramoto (2002, *apud* MAIA; SOUZA, 2010), as palavras apresentam vários problemas em suas propriedades linguísticas, fazendo com que não sejam consideradas boas representantes temáticas. Alguns problemas:

- polissemia: uma palavra com vários significados;
- sinonímia: duas palavras distintas com o mesmo significado;
- duas palavras combinando-se em ordens diferentes produzindo ideias completamente distintas.

Esses fatores podem influir diretamente no resultado de uma busca em SRI (Sistema de Recuperação da Informação) produzindo uma taxa de ruído considerável. E ainda uma baixa precisão e revocação

Segundo Kuramoto (2002, *apud* SILVA *et al*, 2011), a organização da informação baseada em sintagmas nominais permitiria a "navegação na estrutura hierárquica em árvore dos sintagmas nominais", podendo dividí-los em níveis distintos.

Através disso, muitas pesquisas são realizadas para que se possa ampliar o processamento da linguagem natural, identificando o significado expresso em suas estruturas semânticas. (SOUZA, 2005). Com isso, o sistema funcionaria resumidamente da seguinte forma:

1. o usuário fornece o termo que representa o centro do sintagma nominal, ou seja, a palavra central;
2. o sistema irá recuperar todos os sintagmas que tenham essa palavra central, cabendo ao usuário, selecionar o sintagma de segundo nível em sua estrutura;
3. em seguida, o sistema apresenta todos os sintagmas nominais do segundo nível, e isso continua, até que o usuário encontre o sintagma que atenda a sua necessidade de informação.

Nota-se então que a aplicação de métodos automatizados de extração e indexação pelo uso de sintagmas nominais mostra-se bastante promissora, conseguindo conter erros que não seriam possíveis ao se fazer apenas o uso de palavras. (SILVA *et. al*, 2011).

3.8.2 Semântica

A semântica é a disciplina da Linguística que tem por objeto de estudo a descrição das significações próprias às línguas e sua organização teórica. De modo geral, pode-se dizer que a semântica estuda o sentido das coisas. (MECZ, 2006).

Com isso, ela se encarrega do significado, do sentido da frase, enquanto a sintaxe, por sua vez, “determina a forma correta de construção das frases de uma determinada língua, levando em consideração a sequencia de sujeitos, verbos, objetos [...], etc.” (BORGES, 2009). Segundo Baranow (1983) não é possível resolver problemas sintáticos sem recorrer à semântica.

A estrutura semântica pode ser repleta de gírias, regionalismos, jargões, todo o universo de palavras que as pessoas de uma determinada língua têm à sua disposição para expressar-se, constituindo o chamado léxico. Essa estrutura lexical contem um conjunto de vocábulos de uma determinada língua, abrangendo o conhecimento linguístico de uma dada sociedade, de modo que essa mesma estrutura possui valor diferente de uma língua para outra.

A semântica se forma com diversas “teorias metodológicas oriundas da linguística e que partem de uma base semântica do estudo das línguas” (CAFÉ; BRASCHER, 2011). A seguir serão apresentadas de forma sucinta algumas dessas teorias:

Teoria de Valência

Desenvolvida por Tesniere em 1959, afirma que a frase é um conjunto organizado formado por palavras e pelas conexões estabelecidas entre as mesmas. Segundo Tesniere (1959, *apud* CAFÉ; BRASCHER, 2011) essa conexão é indispensável para a compreensão e expressão do pensamento. Borba (1996 *apud* CAFÉ; BRASCHER, 2011) aponta a existência de uma gramática de valência, a qual seria composta por três elementos básicos: argumento, predicado, e relação entre eles; abrangendo a dimensão sintática e a dimensão semântica.

Gramática de casos

Segundo Filmore (1968, *apud* CAFÉ; BRASCHER, 2011) desenvolveu um modelo de gramatica que considera que a sentença, de forma básica, é formada por um verbo e um ou mais sintagmas nominais, onde cada um deles está associado ao verbo.

Gráficos conceituais

Surgiu em 1968 com Sowa, por meio da aplicação da ideia de fluxogramas para criar um modelo que representasse o conhecimento utilizando caixas e círculos para criar gráficos conceituais. (CAFÉ; BRASCHER, 2011)

Os gráficos conceituais constroem uma linguagem de representação do conhecimento, e é constituído da seguinte maneira, (CAFÉ; BRASCHER, 2011):

a) Os conceitos: são os conteúdos de pensamento; representam entidades, ações.
b) As relações: apresentam as ligações existentes entre conceitos e os papéis de cada entidade.

Teoria da gramática funcional de Simon Dik

Oferece uma base consistente para interpretar as expressões linguísticas e também para sintetizar a análise. Por meio dela, é possível organizar a informação veiculada pela expressão linguística de modo que seja possível automatizar o estudo da língua geral ou especializada. (CAFÉ; BRASCHER, 2011)

3.9 Experimentos de processamento de linguagem natural

Nesta seção, serão apresentados os softwares encontrados na literatura para processamento de linguagem natural, alguns em funcionamento, outros ainda em fase de testes e, também, alguns protótipos. Cabe ressaltar que

alguns softwares serão apresentados de maneira breve, enquanto outros de forma mais extensiva.

A fim de facilitar a compreensão, optou-se por apresentar as características dos softwares levantados de forma similar àquela desenvolvida por Pérez, Alfonseca e Rodríguez (2005). Os autores consolidaram os dados descritivos de softwares de indexação automática, levantados na literatura entre 1966 e 2004, na da tabela abaixo (Tabela 1):

Tabela 1 - Softwares

System	Reference	Technique	Results	Domain
PEG	(Page, 1966)	Statistical	Corr: .87	Non factual disciplines
E-rater	(Burstein, Leacock, & Swartz, 1998)	Statistical and NLP	Agr: .97	Non-native English writing
Larkey's system	(Larkey, 1998)	Text Categorization	EAgr: .55	Social and opinion
IEA Landauer, 1999)	(Foltz, Laham &	LSA	Agr: .85 military	Psychology and
SEAR	(Christie, 1999)	Information Extraction	Corr: .45	History
Apex Assessor	(Dessus, Lemaire & Vernier, 2000)	LSA	Corr: .59	Sociology of education
IEMS	(Ming, Mikhailov & Kuan, 2000)	Indextron	Corr: .8	Non mathematical
IntelliMetric	(Vantage, 2000)	NLP	Agr: .98	k-12 and creative writing
ATM	(Callear, Jerrams-Smith & Soh, 2001)	Information Extraction	*Not Available*	Factual disciplines
C-rater	(Burstein, Leacock, & Swartz, 2001)	NLP	Agr: .83	Comprehension & algebra
Automark	(Mitchell, Russell, Broomhead, & Aldridge, 2002)	Information Extraction	Corr: .95	Science
BETSY	(Rudner & Liang, 2002)	Bayesian networks	CAcc .77	Any text classification
PS-ME	(Mason & Grove-Stephenson, 2002)	NLP	*Not Available*	NCA or GCSE exam
CarmelTC	(Rosé, Roque & VanLehn, 2003)	Machine Learning and Bayes classif.	f-S: .85	Physic
Auto-marking	(Sukkarieh, Pulman & Raikes, 2003)	NLP & Pattern Match.	EAgr: .88	Biology
Atenea	(Alfonseca & Pérez, 2004)	Statistical, NLP and LSA	Corr: .56	Computer Science

Fonte: Pérez; Alfonseca; Rodríguez (2005).

Desta forma, os softwares identificados pelo levantamento bibliográfico deste trabalho também foram consolidados em uma tabela (Tabela 2), considerando os seguintes aspectos: a referência bibliográfica, a técnica e os resultados descritos:

Tabela 2 – Experimentos em Processamento de Linguagem Natural

Sistema	Referência e ano	Técnica	Resultado
SPIRIT	P. Binquet et. Al, 1983	Indexação automática baseada em métodos linguísticos e estatísticos.	Armazenamento e interrogação em linguagem natural.
Hirst	Hirst, 1987.	Usa um analisador sintático, um interpretador semântico.	Lida com ambiguidades léxicas e estruturais.
Automindex	Robredo, 1991	Faz uso de dois antidicionários de palavras vazias.	
Analisador morfossintático	Brito, 1991	Parser	Transpor automaticamente um texto em linguagem natural para uma metalinguagem de análise gramatical.
SRIAC	Kuramoto, 1997	Extração de sintagmas nominais.	Usuários montam sua própria expressão de busca.
IILLICO	Pasero & Sabatier, 1998	Software genérico de linguagem natural.	Executa tarefas linguísticas específicas como análise, síntese e composição.
Kana Customer Messaging System	Scott, 1999	Processamento de linguagem natural.	Categoriza a entrada de emails, encaminha ao departamento correto e agiliza o processo de resposta.

Brightware	Scott, 1999	Utiliza técnicas de PLN.	Prova o significado de grupos de palavras e r eponde emails automa ticamente.
SISA	Gil Leiva, 1999	Linguagem documentária .	Faz uma comparação entre o documento e u ma linguagem docume ntária.
NPLWin	Elworthy, 2000	Faz uso de PLN.	Aceita frases e proporc iona a análise sintática detalhada.
Zstation	Brascher, 2002 .	Tratamento automático d e linguagem natural.	Analisa as frases em s uas propriedades sem ânticas e morfológicas.
Semantic A gent	Lucena, 2003	Processamento de lingua gem natural.	Capaz de compreende r solicitações do usuári o em linguagem natura l.
Thought Tr easure	Erik T Muller, L ucena, 2003	Uso de ontologias e rede s semânticas.	Organiza os conceitos e as relações ontológic as de forma hierárquic a.
Atenea	Pérez, 2005.	Faz uso de PLN e de AS L, dando pontuações aos textos.	Combinando as duas t écnicas, há uma melho ra na pontuação.
SiRILiCO	Gottschalg, 20 05.	Baseia-se em teorias da linguística computacional e ontologias.	hierárquica a partir da l inguística de textos.
DocMir	Behera, 2007.	Usa 3 ferramentas: ferra menta de captura; ferram enta de análise e ferrame nta de recuperação.	Indexa e recupera documentos de reuniões, conferências, seminários.
Indexação	Câmara Júnior	Processamento de lingua	Indexação automática

automática de acórdãos	, 2007	gem natural.	dos textos.
Indexação automática de vídeos	Pimentel Filho, 2008	Sumarização e indexação automática de vídeos digitais.	Visa oferecer suporte as operações de busca de conteúdos visuais.
Coh – Metrix e LIWC	Duran et. Al, 2009	Processamento de linguagem natural.	Avalia a conversação entre dois participantes interativos.
EMR	Gilles, 2013.	Faz uso de taxonomia hierárquica.	Indexação mais profunda, precisa e ágil.

Ladeira (2010) em sua tese de doutorado elaborou um estudo acerca da produção brasileira na área de PLN, abarcando os anos de 1973 a 2009. Como resultado, a autora apresentou um mapa conceitual com as principais problemáticas da área de PLN identificadas pelo seu estudo. Importante ressaltar que não cabe abordar cada uma das problemáticas apresentadas pela autora, apenas mencioná-las como mais uma fonte de conhecimento.

As problemáticas encontradas foram organizadas em um mapa conceitual e também sucintamente em uma tabela, apresentada a seguir:

Tabela 3 - Principais problemáticas reveladas apartir da análise de conteúdo

Principais problemáticas reveladas a partir da análise de conteúdo

Problemática	#Artigos relacionados
Recuperação de informação	18
Sumarização	10
Tratamento de Ambiguidade	10
Analisadores (*parser*)	10
Tradução	9
Aplicações para a própria área	4
Exemplos de aplicações do PLN	4
Correção automática	3
Total	**68**

Fonte: Ladeira (2010)

Devido ao fato de o *parser* estar entre os experimentos abarcados neste trabalho, optou-se por realçar suas problemáticas apresentadas no mapa conceitual de Ladeira (2010). No mapa conceitual mencionado, a autora divide o *parser* em três níveis de análise: análise léxico-morfológica, análise sintática, e análise semântica, e em cada um cita problemáticas apresentadas pelos autores selecionados.

3.9.1 Descrição dos experimentos selecionados

3.9.1.1 Hirst

Em 1987, o sistema produzido por Graeme Hirst foi apresentado em um livro e revisado por Karen Sparck Jones, o sistema apresenta uma abordagem para a linguagem natural, por meios teóricos e práticos. O autor apresenta detalhes e fornece excelentes resumos, fazendo as propriedades de seu trabalho serem bem definidas.

A ideia de Hirst era construir um sistema de interpretação que pudesse resolver ambiguidades léxicas e estruturais. Sua preocupação é

essencialmente computacional, ele não faz reclamações sobre a psicolinguística, contudo ele é disposto a explorar 'psicolinguisticamente' novas estratégias.

O sistema consiste em um analisador sintático, um interpretador semântico, e dois processadores de desambiguação: o Polaroid Word (PW) subsistema para desambiguação léxica e o Semantic Enquiry Desk para desambiguação estrutural. (SPARCK JONES, 1987).

3.9.1.2 ILLICO

É um software genérico de linguagem natural construído para executar tarefas linguísticas específicas, como a análise, síntese e composição de sentenças. (PASERO; SABATIER, 1998, *apud* CHOUDHURY, 2003).

3.9.1.3 Kana Customer Messaging System

Ele pode categorizar a entrada de e-mails, encaminhá-los ao departamento certo e agilizar o processo de resposta. Também tem uma função de auto-sugestão que ajuda um cliente representante de serviço a responder em território desconhecido. (CHOUDHURY, 2003 apud SCOTT, 1999)

3.9.1.4 Brightware

Sistema que utiliza técnicas de PLN para provar o significado de grupos de palavras ou frases, e responder e-mails automaticamente. (SCOTT, 1999 *apud*, CHOUDHURY, 2003)

3.9.1.5 *NPLWin*

É também um sistema da Microsoft que faz uso de PLN, e aceita frases e proporciona a análise sintática detalhada, juntamente com uma forma lógica (ELWORTHY, 2000 *apud* CHOUDHURY, 2003).

3.9.1.6 DocMIR

Apresentado por Behera, Lalanne e Ingold (2007), é um sistema automático que suporta documentos de reuniões, indexando e recuperando. A arquitetura do sistema foi desenvolvida para capturar, indexar automaticamente, e recuperar reuniões, conferências, seminários e etc. Ele é formado por três ferramentas principais:

- uma ferramenta de captura: ela permite que os dados das reuniões sejam capturados e arquivados. Nessa ferramenta, os slides são sincronizados automaticamente sem ser preciso instalar nenhum software no computador. Essa captura é feita por meio de câmeras, webcams, câmeras com microfones para capturar as conversas, entre outros meios;
- uma ferramenta de análise e indexação: os vídeos capturados são usados por essa ferramenta para a realização da indexação automática;
- uma ferramenta de recuperação: é uma ferramenta interativa, que se beneficia das palavras-chave e/ou dos documentos capturados dos dispositivos portáteis para acessar os vídeos arquivados. Depende da segmentação dos métodos usados, das performances correspondentes e da qualidade da indexação.

3.9.1.7 EMR

A sigla faz referencia ao *Eletronic Medical Records*, que são os prontuários médicos eletrônicos. Acessar essa informação de forma segura e também torná-la acessível para pesquisa, depende de uma padronização taxonômica para organizar e indexar o conteúdo. (GILLES, 2013)

Vocabulários controlados são necessários para interpretar o conteúdo. As quase-taxonomias fornecem códigos, contudo esses códigos se tornam difíceis no momento da indexação, exigindo uma tradução da linguagem natural da EMR para uma equivalente. (GILLES, 2013)

Segundo o mesmo autor, usar um sistema de categorização com uma taxonomia hierárquica permite uma indexação mais profunda, precisa e uma filtragem rápida e automática dos conceitos mais gerais.

Desse modo, Gilles (2013) faz um estudo e apresenta algumas metodologias utilizadas na classificação automática e semi-automática no campo da medicina. Como resultado, ele obtém que muitos sistemas utilizam uma mistura de métodos para alcançar o resultado desejado.

A maioria dos sistemas requerem uma taxonomia em ordem para começar, e muitos sistemas também colocam '*tags'* nos textos para cada palavra chave na taxonomia. Desse modo a taxonomia permite uma melhor indexação e filtragem. (GILLES, 2013)

Contudo, em meio a tantas diferenças de organização dos sistemas, o fundamental é trazer clareza e precisão para a linguagem, ou seja, superar a distância entre a pergunta do usuário e a resposta dos sistemas.

3.9.1.8 Coh- Metrix e LIWC

O Coh-Metrix é uma ferramenta para o processamento da linguagem natural, foi utilizado em um estudo pela Universidade de Memphis, apresentado por Duran et. al (2009) para avaliar a comunicação 'verdadeira' ou 'falsa' que ocorre dentro do computador. Nesse estudo, o Coh-Metrix é comparado com uma outra ferramenta de PLN, o chamado Linguistic Inquiry and Word Count (LIWC).

Ambos trabalham em cima da conversação entre dois participantes interativos, ou seja, mensagem instantânea, por meio dela, não acontece o encontro ao vivo entre os integrantes, o que facilita uma falsa conversa, chamada também de 'enganosa'.

Desse modo, o LIWC é uma ferramenta que avalia mais de 70 dimensões da linguagem, de modo que ele rastreia características linguísticas como indicativos de fenômenos sociais e psicológicos, entre eles personalidade, expressão emocional e saúde mental.

Segundo Duran *et. al* (2009), comparando as duas ferramentas citadas anteriormente, pode-se oferecer uma análise única e mais completa acerca da natureza desse tipo de linguagem dita pelo autor como 'enganosa'.

3.9.1.9 Semantic Agent

O Semantic Agent é um protótipo de uma plataforma que visa o desenvolvimento de agentes “que sejam capazes de compreender solicitações em linguagem natural, manipular conhecimento e executar ações”. (LUCENA, 2003).

O protótipo visa a criação de um agente de software capaz de realizar tarefas de interesse do usuário de maneira automática, sem que o mesmo precise atuar diretamente na execução da tarefa.

3.9.1.10 Thought Treasure

É um software para processamento de linguagem natural que se baseia em ontologias, aplicando os conceitos de rede semântica. O software foi desenvolvido por Erik T Mueller, o programa mantem uma ontologia com informações sobre senso comum.

Essa ontologia é composta por conceitos e asserções que estabelecem relações entre conceitos. A base de conceitos é composta por 27.093 conceitos e 51.305 asserções. O software organiza os conceitos e as relações ontológicas de forma hierárquica. (LUCENA, 2003)

3.9.1.11 SPIRIT

É um sistema de indexação automática baseado em métodos linguísticos e estatísticos, possui o objetivo de processar os documentos em linguagem natural. Foi desenvolvido por P Binquet, F. Debili, C. Fluhr e B Pouderoux do Centre National de la Recherche Scientifique (CNRS).

Segundo Andreewski e Ruas (1983), o programa permite o armazenamento e a interrogação em linguagem natural; conta ainda com o tratamento linguístico a todos os níveis textuais introduzidos no sistema, juntamente com tratamentos estatísticos; permitindo a realização de uma indexação ponderada dos documentos.

Desse modo, segundo os mesmos autores, quando uma pergunta é feita ao sistema, tudo em linguagem natural, os documentos obtidos como resposta são classificados de acordo com um critério de proximidade semântica. A

seguir serão apresentados, de forma sucinta, os componentes do sistema, visto que não tem-se a pretensão de aprofundamento no funcionamento do programa.

Os componentes são os seguintes:

a) um dicionário, que permite a análise morfológica dos textos;
b) algoritmos de análise sintática, corrigindo ambiguidades;
c) algoritmos de análise semântica, fazendo a identificação correta da relação palavra em função do contexto.

3.9.1.12 Sistema de Indización Semi-Automático (SISA)

É um software de indexação semi-automática proposto pelo Prof. Dr. Isidoro Gil Leiva da Universidade de Murcia na Espanha (1999, 2008), decorre de um estudo sobre automatização da indexação. Foi desenvolvido inicialmente para a área de biblioteconomia e documentação, contudo, permite adaptar sua configuração para atuar em qualquer área do conhecimento. (NARUKAWA; GIL LEIVA; FUJITA, 2009)

Segundo os mesmos autores, seu processamento se desenvolve em três módulos, mas de maneira geral o software faz uma comparação entre o documento – título, resumo e texto – e uma linguagem documentária, utilizando critérios de frequência determinados pelo software para indicar os termos de indexação.

Além disso, segundo Lima e Boccato (2009), o software relaciona em uma lista à parte os termos que são candidatos a descritores, ou seja, aquelas palavras que não estão na linguagem documentária, mas que ocorrem várias vezes no texto indexado.

3.9.1.13 Atenea

É um sistema capaz de fazer perguntas de forma aleatória conforme o perfil dos alunos, fazendo um método de pergunta-resposta para os estudantes e então atribuindo uma pontuação. O sistema funciona em inglês e espanhol.

Os resultados provam que para todos os conjuntos de dados, quando técnicas de PLN são combinadas com técnicas de ASL (análise semântica latente), a correlação entre as pontuações dadas por Atenea e a pontuação

dada pelos professores para o mesmo conjunto de dados, melhora. (PÉREZ, 2005)

O autor afirma que isso se deve à complementaridade entre ASL, onde as palavras são tratadas mais em um nível semântico, e as técnicas de PLN usadas em Atenea, são mais focadas no nível lexical e sintático.

3.9.1.14 Zstation

Software apresentado por Brascher (2002) em seu estudo sobre ambiguidades. É um sistema de tratamento automático da linguagem natural, onde seu ponto inicial é que para desempenhar uma tarefa, como por exemplo fazer a análise de uma frase, é necessário coletar toda informação sobre esta frase; tanto relativo a propriedades semânticas e morfológicas das palavras e frases, e suas possíveis conexões, fazendo com que o conhecimento coletado permita uma ou várias interpretações.

Ele funciona em módulos que envolvem: morfossintática, sintagmática e semântica. Faz uso de um dicionário automático, uma gramática morfológica e uma gramática de argumentos, levando em conta ainda os conceitos de ontologia. Mostra-se como um sistema eficaz mas que possui grande complexidade de implementação e manutenção (BRASCHER, 2002).

3.9.1.15 SRIAC

É uma proposta para um novo sistema de recuperação de informação (SRI), baseado em sete princípios para orientar o sistema: leveza, precisão, rapidez, visibilidade, solidez, interatividade. Segundo Kuramoto (1997), nesse sistema “o procedimento de indexação proposto consiste na extração dos sintagmas nominais e na sua indexação, como descritor, segundo uma estrutura em árvore.”

Pretende-se obter como resultado um SRI que ofereça aos usuários a oportunidade de montar a sua própria expressão de busca, de maneira indireta, por meio da navegação na estrutura dos sintagmas nominais até o momento que o usuário encontre o sintagma que satisfaça sua necessidade de informação (KURAMOTO, 1997).

3.9.2 Experimentos brasileiros

3.9.2.1 SiRILiCO

É uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologias apresentado por Duque-Gottschalg em 2005. O protótipo faz uso de programas já desenvolvidos e disponibilizados para uso, como o programa *Palavras*, e o programa *Protegé*. Contudo, foi desenvolvido um software específico para o analisador semântico, chamado de *GeraOnto*.

Cabe aqui explicar do que se trata cada programa, o *Palavras* é um analisador sintático gratuito na web, que usa regras gramaticais baseadas na *Constraint Grammar Formarlism*; já o *GeraOnto* é um analisador semântico que gera uma 'ontologia leve',e foi desenvolvido a partir de outro, o *SMOSe*.

O Protegé foi desenvolvido na Universidade de Stanford e é um editor de ontologias também gratuito, feito em Java. Vale ressaltar que ele permite modificações por parte dos usuários, pois se trata de um *open source*.

O SiRILiCO é na verdade composto de vários módulos, os quais serão apresentados de forma sucinta para a melhor compreensão do sistema.

A utilização do Módulo de processamento de linguagem natural (MPLN) é feita para otimizar a indexação, de modo que os textos serão indexados em função dos conceitos, analisando as frases no documento.

A atomização do texto, é a divisão do texto em partes, onde o autor, título, e as palavras-chave são enviados para o SMOF; já as frases que compõem o texto são enviadas para o SMOSi para serem processadas sintaticamente.

O SMOSi é responsável por processar sintaticamente cada frase do texto, depois disso, o produto será enviado para o SMOSe, onde ocorre a análise semântica.

Após a realização da etiquetagem sintática, os elementos semânticos são identificados e por sua vez, descriminados. Parte-se então para o Módulo Gerador de Ontologia (MGO), fazendo uso do Protegé, de acordo com Duque-Gottschalg (2005): " os conceitos extraídos dos textos da coleção tornam-se então as classes da ontologia gerada pela coleção".

Segue então o Sub-Módulo de Ontologia Básica (SMOB), que se trata de uma ontologia criada e armazenada no Protegé. É o padrão de referência para converter automaticamente as etiquetas sintáticas em etiquetas semânticas.

O Sub-Módulo de Ontologia Formada (SMOF) se refere a uma ontologia dita ‘leve’, criada automaticamente dos conceitos encontrados nos textos da coleção e mantida no Protegé. Vale ressaltar que essa ontologia serve como base para a geração do índice da coleção.

O Módulo Gerador de Índice (MGI), é o responsável pela edição de ontologia no Protegé. Há ainda o SMEI, que funciona como uma lista invertida de conceitos, onde para cada conceito existe uma lista com os textos nos quais aqueles mesmos conceitos aparecem.

Segundo o autor, foi realizado um experimento-piloto e um experimento de validação, de modo que em seu experimento-piloto o SiRILiCO utilizou 41 artigos, em língua portuguesa, todos publicados na Revista Ciência da Informação, revista 31, números 1, 2 e 3 de 2002, e revista 32, número 1 de 2003; já no experimento de validação foram acrescentados mais 180 artigos, contudo, foram utilizados os mesmos módulos do experimento-piloto.

Para a realização dos experimentos, Duque-Gottschalg (2005) afirma que foram utilizados apenas o título, autor, palavras-chave e a introdução dos referidos artigos.

O autor conclui afirmando que a linguística computacional e a ontologia podem oferecer grandes contribuições para a recuperação e disseminação da informação, especialmente no que diz respeito ao tratamento automático. Após os estudos do protótipo é possível dizer que os usuários podem beneficiar-se significativamente de uma estrutura hierárquica desenvolvida a partir da linguística dos textos.

Por fim, Duque-Gottschalg (2005) conclui que: “é viável a criação de uma ontologia leve automaticamente única e exclusivamente a partir de análises sintáticas e semânticas dos textos da coleção da qual se quer uma ontologia”.

Além disso, o sistema contribuiu de forma clara para área da Ciência da Informação, mostrando que é possível desenvolver um modelo de recuperação de informação, fazendo uso de teorias de Linguística. Gerando então, uma

ampla possibilidade de estudos, como na geração automática de índices; recuperação automática da informação e utilização de ontologias para busca do usuário.

3.9.2.2 Automindex

É um sistema de indexação apresentado por Robredo (1991). Possui como característica principal a existência de dois antidicionários concomitantes de palavras vazias: um de palavras invariáveis, e outro de raízes de palavras não significativas para uma determinada área do conhecimento. (NARUKAWA; GIL LEIVA; FUJITA, 2009).

Segundo Robredo (1991), para o processamento são levados em conta os títulos e os resumos. O software funciona da seguinte maneira: primeiro o texto é analisado comparando as palavras do texto com as palavras do dicionário de invariáveis, caso constem nesse dicionário, serão desprezadas.

O mesmo processo é feito com as palavras comparadas com o dicionários de raízes significativas. As palavras que restarem serão os possíveis descritores. No estágio final, para serem de fato selecionadas como descritores, as palavras são comparadas com um dicionário de palavras significativas, caso constem nesse dicionário, serão descritores, caso contrário, serão ‘possíveis’ descritores (NARUKAWA; GIL LEIVA; FUJITA, 2009).

3.9.2.3 Analisador morfosintático

Consiste em um estudo realizado por Brito (1992), onde o autor realiza um teste com as Gramáticas Afixo. Desse modo pretende “transpor automaticamente um texto, em linguagem natural, para uma metalinguagem de análise gramatical” . Essa análise será capaz de mostrar a ordem estrutural dos constituintes da frase e também a sua ordem linear, indicando as dependências que ocorrem entre os elementos do enunciado.

Segundo Brito (1992), deve-se levar em conta a variedade dos componentes que intervém na linguagem: morfológicos, lexicais, sintáticos, semânticos, lógicos. Desse modo, ele apresenta uma visão diferente, com uma descrição mais rica e elaborada dos fenômenos linguísticos e que auxiliam de forma direta as ideias sobre o tratamento automático da informação.

Com isso, o autor indica o uso de Gramáticas Afixos, mostrando que “pode-se aumentar a qualidade dos resultados das análises morfossintáticas por meio de uma descrição gramatical mais bem adaptada, mais fina e mais fiel ao modelo linguístico proposto”. (BRITO, 1992)

Como resultado procurou-se uma indexação mais uniforme. Houveram alterações na gramática de análise, obtendo-se um parser melhorado e a ideia de um sintagma nominal como descritor.

Desse modo, ele espera que seu estudo sirva de instrumento para uma nova geração de sistemas de recuperação da informação com uso de sistemas de indexação automática, auxiliando na melhora e na evolução da área.

3.9.2.4 Programa de indexação de vídeos

Pimentel Filho (2008) propõe um ambiente para sumarização e indexação automática de vídeos digitais com o intuito de oferecer suporte as operações de busca baseadas em conteúdo visual e em repositórios de vídeo. Afirma que " é possível se obter uma representação bastante reduzida do vídeo através de quadros-chave, que armazenam informações suficientes sobre as características visuais do conteúdo do vídeo.” (PIMENTEL FILHO, 2008)

O autor apresenta em seu trabalho de mestrado um estudo sobre um ambiente de indexação e recuperação de conteúdo de vídeos. Explica que a arquitetura do ambiente foi dividida em dois módulos principais: o vídeo parsing e o vídeo oráculo.

O vídeo parsing é o responsável pelo parsing do fluxo do vídeo, ou seja, fará a separação do vídeo em quadros. Já o outro módulo, chamado vídeo oráculo é responsável pela indexação, recuperação e navegação; consumindo os dados produzidos pelo vídeo parsing e alimentando um banco de dados com o resultado do processamento.

3.9.2.5 Indexação automática de acórdãos

Câmara Júnior (2007), realizou um estudo com o objetivo de indexar automaticamente documentos de acórdãos. O autor recolheu acórdãos de direito penal da base de jurisprudência do Tribunal de Justiça do Distrito Federal e Territórios do período de 1997 a 2007.

Um dos instrumentos utilizados foi um analisador denominado Qtag, que tem como objetivo realizar o processamento de linguagem natural dos textos selecionados. É um software livre que funciona ainda como um etiquetador probabilístico morfológico construído para qualquer idioma. (CÂMARA JÚNIOR, 2007)

Foi utilizado ainda, um sistema de análise, desenvolvido e utilizado para extração de estruturas a partir do texto analisado pelo Qtag, montando estruturas complexas em formatos definidos. O autor relata que por fim, foi utilizada uma ferramenta de atribuição de índice baseado no tesauro jurídico para realizar o final do processo.

O tesauro, segundo Câmara Júnior (2007), surge como um meio de oferecer mecanismos que possam aprimorar os índices de revocação e precisão nas pesquisas, através das relações que apresenta.

3.10 Aplicações do PLN na internet, na web e nas bibliotecas digitais

A internet e a web têm trazido melhorias significativas na maneira em que as pessoas criam, olham, e usam a informação. Um grande volume de informações está agora disponível através da internet e das bibliotecas digitais. Com isso, esse desenvolvimento tem trazido alguns problemas relacionados com o processamento da informação e a sua recuperação.

Dentro desse contexto, Neves (2009) afirma que com o passar do tempo, houve um volume crescente de informação disponibilizada na internet, de modo que o tratamento desse volume de informação necessita de mecanismos que possam otimizar a sua execução.

Conforme Rocha (2004), com a super oferta de informação proporcionada pela internet, as pessoas acabam tendo dificuldade em encontrar as informações que lhe são relevantes. Isso se deve, de acordo com o mesmo autor, ao fato de que existe pouca organização da informação na web, impedindo estratégias e mecanismos de busca que funcionem eficientemente.

Segundo Ramalho, Vidotti e Fujita (2007), o avanço exponencial na quantidade de recursos informacionais que estão disponíveis no ambiente web

mostra que os modelos clássicos de representação e recuperação da informação precisam ser revistos e repensados sob diferentes perspectivas.

Neves (2009) apresenta ainda o fato de que na internet, portadora de repositórios de informações, não dispõe de profissionais de referência, como aqueles encontrados em uma biblioteca física. Portanto, o papel do bibliotecário de referência é desempenhado pela ferramenta de busca da web ou até mesmo pelo próprio usuário.

Pode-se afimar então, que a indexação na web é extremamente necessária para a boa recuperação da informação por parte de seus usuários. Essa indexação é formada por quatro elementos inter-relacionados, apresentados a seguir:

> - " Metadados: tem a função de descrever e ordenar a informação no documento;
> - Posicionamento web: é o ranqueamento das pesquisas, sendo realizado pelas ferramentas de busca;
> - Buscadores: diretório ou uma ferramenta de busca. O diretório é organizado manualmente, já a ferramenta de busca é organizada por meio de um algoritmo que leva em consideração a relevância;
> - Usuário: visto como um documentalista, pois recorre a internet para a sua busca de informações, contudo já está familiarizado com conceitos da área de documentação." (GIL LEIVA, 2007, apud NEVES, 2009)

Com todos esses pontos, a organização dos documentos na web, e a recuperação da informação, não são tarefa fácil, de modo que a subjetividade vem intensificada nesse meio, pois a visão do usuário é comumente deixada de lado.

A indexação surge então como um meio de auxiliar na precisão da busca por essa informação. Segundo Souza (2000, *apud* ROCHA, 2004), a internet não faz a seleção de nenhum tipo de documento (ao contrário de uma biblioteca tradicional), de modo que abrange todas as áreas do conhecimento e torna como seus usuários, todas as pessoas que a acessam, não diferenciando por tipo.

Outro problema, segundo Rocha (2004), diz respeito ao fato de que muitas informações da internet são irrelevantes e dúbias, sendo necessária uma filtragem por parte do usuário, contudo, segundo o mesmo autor, a maioria dos usuários não tem uma clareza com relação ao que deseja obter em suas buscas.

A indexação automática seria uma boa solução, contudo, ela também não consegue alcançar o ponto do usuário, uma vez que ela apenas teria como fonte o texto digitalizado. Desse modo, a indexação automática na web, é necessária, mas precisa ainda de muitos estudos. Segundo Neves (2009), a fase atual dos sistemas de indexação automática é marcada pela união do processamento da linguagem natural e dos sistemas inteligentes (sistemas apoiados na inteligência artificial).

De acordo com algumas pesquisas apresentadas por Choudhury (2003), cerca de 80% das fontes de informação da internet e bibliotecas digitais disponíveis são atualmente em inglês. Isso faz com que seja necessário o estabelecimento de sistemas multilíngues de informação.

Partindo desse ponto, várias abordagens têm sido propostas para a tradução, como dicionários bilíngues para converter termos de uma linguagem fonte para uma linguagem alvo, mas isso pode acabar sendo um problema também, pois muitas vezes o dicionário não traduz a palavra corretamente de acordo com o sentido expresso.

Staab et. al. (1999) descreve as características de um agente de informação inteligente chamado GETESS, o qual usa métodos semânticos e capacidades de PLN para reunir informações turísticas da web e apresentá-las ao usuário humano.

Devido ao volume de texto disponível na internet, muitos pesquisadores têm proposto utilizar a web como teste para pesquisas em PLN, pois apesar dos ‘ruídos’ o texto da web apresenta a língua da forma que ela é usada, e as estatísticas derivadas da web podem ter usos práticos na área de PLN. (GREFENSTETTE, 1999 apud CHOUDHURY, 2003)

Cabe ressaltar que o PLN, também trabalha com a avaliação, pois esta é uma área importante em qualquer sistema. Os pesquisadores da área de Ciência da Informação têm desenvolvido métodos confiáveis de avaliação nos sistemas de PLN.

Um desses programas criados é o ELSE (*Evaluation in Language and Speech Engineering*), que se refere a um projeto da Comissão Europeia e teve como objetivo estudar a possível implementação de avaliação comparativa em sistema de PLN; essa avaliação consiste em um conjunto de participantes que comparam os resultados dos seus sistemas, que fazem uso de tarefas

semelhantes e dados relacionados. O ELSE consórcio identificou 5 tipos de avaliação:

- avaliação de pesquisa básica: tenta validar uma nova ideia ou estimar as melhorias que foram trazidas em relação aos métodos mais antigos;
- avaliação tecnológica: tenta estimar o desempenho e adequação de uma nova tecnologia para resolver um problema;
- avaliação de uso: tenta estimar a usabilidade da tecnologia para resolver um problema real.
- avaliação de impacto: tenta medir as consequências socioeconômicas da tecnologia;
- avaliação de programa: tenta determinar quanto vale a pena financiar um programa para uma dada tecnologia

EAGLES (*The Expert Advisory Group on Language Engineering Standards – Evaluation Workgroup*), fase 1, e o fase 2: EAGLES-I, vem de uma iniciativa europeia que propõe uma avaliação centrada no usuário do sistemas de PLN. O trabalho do EAGLES toma como ponto de partida um padrão existente, a ISO 9126, a qual está relacionada principalmente com a definição das características de qualidade para serem usados na avaliação dos produtos de software.

Importante ressaltar que segundo Choudhury (2003), o MUC, *Message Understanding Conferences*, que por sua vez não existe mais, foi o pioneiro em abrir uma plataforma internacional para compartilhar pesquisas na área de sistemas de PLN.

3.11 Recuperação da informação

A recuperação da informação é uma das áreas de aplicação do processamento de linguagem natural, visando uma recuperação eficiente em todos os pontos. Contudo, é um grande desafio fazer a tecnologia de PLN funcionar de forma eficaz e eficiente, e também realizar testes de avaliação adequados para concluir em que medida a abordagem funciona em um ambiente de pesquisa interativa.

Segundo Cunha (1999), existem muitas pesquisas na área de bibliotecas digitais para desenvolver técnicas de indexação que independam de uma representação textual. Além disso, o autor afirma que existem novos tipos de documentos que são incorporados ao acervo de uma biblioteca, como por exemplo video conferências. Com isso, são necessárias novas formas de indexação como um meio de atribuir pontos de acesso a esse tipo de documento.

Conforme apresentado por Ramalho, Vidotti e Fujita (2007), uma das preocupações dos pesquisadores em indexação é a evolução rápida de técnicas de recuperação automática de informação, promovendo o aumento da responsabilidade do indexador ao determinar o assunto do documento. Essas novas formas de recuperação da informação exigem um maior aprofundamento teórico do indexador, evitando o risco de uma prática descompromissada com a representação do conteúdo do documento.

Estão inseridos nesse contexto os SRI, Sistema de Recuperação da Informação, que seriam os responsáveis por armazenar os dados, distinguindo as informações que foram armazenadas por um usuário, das que serão obtidas por outro. (SOUZA, 2005).

Segundo Lancaster e Warner (1993, *apud* SOUZA, 2005), os SRIs são uma interface entre recursos de informação, seja em meio impresso ou digital, e uma população de usuários, desempenhando tarefas como: aquisição e armazenamento dos documentos, organização e controle, distribuição e disseminação aos usuários.

Desse modo, Souza (2005) afirma que os SRIs têm a função de organizar e viabilizar o acesso aos itens de informação, realizando as seguintes atividades:

- representação das informações que um documento contém, por meio da indexação e da descrição dos documentos;
- armazenamento e gestão física desses documentos;
- recuperação das informações e dos próprios documentos que foram armazenados, satisfazendo os usuários e suas necessidades de informação;

Vale ressaltar que para o sucesso do PLN, as técnicas devem aplicadas em conjunto com outras tecnologias, como a visualização, reconhecimento de

voz e agentes inteligentes.(FELDMAN, 1999, *apud* CHOUDHURY, 2003). Algumas vezes, deve-se realizar a aplicação de técnicas diferentes para consultas diferentes, alguns resultados serão melhores se usados pesquisadores booleanos, contudo, outros serão mais eficientes se usados a linguagem natural.

Zadrosny et al.(2000, *apud* CHOUDHURY, 2003), sugere que em um ambiente ideal de recuperação de informações, os usuários devem ser capazes de expressar seus interesses ou consultas de forma direta e natural, seja falando ou escrevendo; e o sistema de computador, por sua vez, deve então ser capaz de fornecer respostas inteligentes as perguntas.

Contudo, apesar de muitos estudos, essas metas não podem ser plenamente alcançadas, devido às limitações da ciência, tecnologia, conhecimento e ambientes de programação. Entre os principais problemas estão:

- limitações no entendimento do PLN;
- gerenciamento das complexidades de interação;
- falta de modelos precisos do usuário;

3.12 Folksonomia

De acordo com Guedes, Moura e Dias (2011), em meio ao contexto informacional onde se forma a World Wide Web, surge a importância do pensamento dialógico para estudar os ambientes sociais semânticos que se baseiam em folksonomias.

Segundo os mesmos autores, são espaços colaborativos onde há uma grande troca e mediação de informações, além da geração de diversos conhecimentos pelos usuários que interagem naquele espaço.

Folksonomia é definido por Vander Wal (2007, *apud* GUEDES, MOURA, DIAS, 2011), como:

> (..) o resultado da livre marcação pessoal de informações e objetos para uma recuperação do mesmo. A marcação é feita em um ambiente social. A folksonomia é criada a partir do ato de marcação pela pessoa que consome a informação.

Com isso, a folksonomia estuda a organização da informação a partir da visão de seus usuários, realizando uma representação dinâmica, algo similar

com a mente humana. Por meio do uso de tags, seria uma auto indexação por parte do autor.

Contudo, alguns problemas são apresentados por Guedes, Moura e Dias (2011), como a falta de controle de vocabulário, pois o usuário simplesmente escolhe os termos que fazem mais sentido para ele. Outros problemas como sinônimo, polissemia e inflexão de palavras também acontecem; a falta de hierarquia é outro ponto negativo, pois todas as tags possuem o mesmo valor e se encontram em um mesmo nível. Todos esses aspectos podem influenciar negativamente na recuperação da informação.

Apesar das problemáticas citadas acima, a folksonomia traz mais benefícios se usada corretamente. Desse modo, a indexação em um ambiente folksonômico é mais comumente conhecida como indexação social, que seria aquela orientada pelo usuário. (GUEDES, MOURA, DIAS, 2011)

Segundo Hassan-Montero (2006, apud GUEDES, MOURA, DIAS, 2011), se trata de um novo modelo, onde os próprios usuários dos recursos realizam a descrição de conteúdo, que seria obtida por agregação, de modo que um mesmo recurso seria indexado por vários usuários, obtendo-se uma descrição mais fiel. Trata-se de uma indexação realizada com o uso de linguagem natural e orientada de acordo com as necessidades dos sujeitos que manipulam esses recursos.

3.13 Web semântica

Segundo Bax (2013), a web atual é conhecida como um conjunto de URI's ou Uniform Resource Identifiers, que refere-se exclusivamente a recursos de informação, documentos. Onde um recurso é definido por Berners-Lee et al. (2005 *apud* BAX, 2013) como qualquer coisa que pode ser identificada por um URI.

Importante ressaltar que alguns autores discutem se há uma diferença entre a web semântica e a web 3.0, segundo Rincón (2012, *apud* KUSTER, HERNANDEZ, 2013), os dois conceitos denominam uma mesma realidade. Os mesmos autores apresentam a web semântica como sendo parte da web 3.0, e uma evolução da web 2.0.

A partir da década de 1990, começaram a surgir pesquisas relacionadas com o desenvolvimento de uma 'nova geração web', a qual possibilitaria a incorporação de ligações semânticas aos recursos informacionais, fazendo com que os computadores pudessem compreendê-las de forma automatizada. (RAMALHO; VIDOTTI; FUJITA, 2007).

De acordo com os mesmos autores, Berners-Lee foi quem iniciou os primeiros estudos relacionados com a web semântica, utilizando a expressão : *Machine understandable information*. Segundo Berners- Lee (1999, *apud* RAMALHO; VIDOTTI; FUJITA, 2007), o passo inicial para o desenvolvimento e implantação da web semântica seria fazer a inclusão de dados em um formato que os sistemas de computadores pudessem compreender naturalmente de forma direta ou indireta.

A web semântica visa desenvolver meios para que as máquinas sirvam aos humanos de maneira mais eficiente, contudo, para que isso seja possível, torna-se necessário construir instrumentos que forneçam sentido lógico e semântico aos computadores (RAMALHO; VIDOTTI; FUJITA, 2007).

Segundo Dziekaniak e Kirinus (2004), enquanto a web atual visava ser entendida apenas pelos usuários, a web semântica visa ser compreendida também pelas máquinas, na forma de agentes inteligentes, capazes de operar eficientemente sobre as informações, entendendo seus significados.

Com isso a web 3.0 ou semântica, marca os princípios para criar uma base de conhecimento e informação semântica e qualitativa. Pretende-se com isso poder atender, de forma mais precisa, as demandas de informação e facilitar a acessibilidade aos conteúdos digitais (KUSTER; HERNÁNDEZ, 2013).

De maneira sucinta, a web semântica:

> (...) visa incorporar semântica às informações. Isso proporcionará não somente aos usuários entenderem as informações como também as máquinas. Ela pretende fornecer estruturas e dar significado semântico ao conteúdo das páginas web, criando um ambiente onde agentes de software e usuários possam trabalhar de forma cooperativa (DZIEKANIAK; KIRINUS, 2004).

Conforme Rocha (2004), a web semântica faz uso de metadados para descrever o significado dos recursos da web, além de agente inteligentes, desenvolvidos através de técnicas de inteligência artificial, que usam essas

descrições para auxiliar os usuários da web na localização e manipulação dos recursos.

A proposta da web semântica é permitir que aplicações combinem e processem dados e informações que estão disponíveis na rede. (BERNERS-LEE, 2001, *apud* BAX, 2013). Assim, a web semântica faz uso de RDF(*Resource Description Framework*), um modelo de dados simples, mas expressivo e extensivo, que representa a informação processável por máquina, e os conjuntos de dados são criados de forma independente um do outro. Esse padrão faz uso de URI para identificar recursos de maneira única e global. (PANSANATO, 2007)

A web semântica surge trazendo uma estrutura e significado que permitam a evolução de uma grande rede de documentos para uma rede de dados, onde a informação teria um significado bem definido, podendo ser processada e interpretada por humanos e computadores (ROCHA, 2004).

Segundo Bax (2013), a web semântica não usa a URI apenas para identificar um recurso de informação, ela identifica também qualquer coisa fora do mundo virtual. Dentro desse contexto a web semântica faz a distinção entre sentido e referência, onde a referência é o objeto em si, e o sentido é a descrição RDF do objeto.

Segundo Brascher (2002), as propostas de inclusão de informação semântica em sistemas de busca aplicam diferentes abordagens, de modo que enfatizam um ou outro aspecto da análise linguística, utilizando diferentes métodos de organização do conhecimento. Contudo, todos convergem para a ideia de aumentar a seletividade e eficiência dos motores de busca.

O projeto web semântica visa a criação e a implantação de padrões tecnológicos que permitam não somente a facilidade nas trocas de informações entre agentes pessoais, mas que também estabeleça uma língua comum para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de modo geral (SOUZA, 2005).

Deve-se criar um ambiente onde os softwares agentes possam realizar tarefas para os usuários, fornecendo uma estrutura que contenha conteúdo significativo na web. (BRASCHER, 2002).

Desse modo, é necessária a padronização de tecnologias, linguagens e metadados, de modo que os usuários da web obedeçam a determinadas

regras comuns sobre como armazenar dados e descrever a informação armazenada, possibilitando que essa informação possa ser consumida por outros usuários, de forma automática e sem ambiguidades.

Em meio a isso, para permitir que máquinas façam uso dos metadados para auxiliar os humanos no uso dos recursos são necessárias técnicas de Inteligência Artificial, onde os instrumentos para descrição semântica são chamados de ontologias (ROCHA, 2004).

Para os computadores poderem ler o conteúdo da web, é necessário que eles consigam ler dados estruturados e tenham acesso a um conjunto de regras para conduzirem o raciocínio. Desse modo, as páginas web terão de ser escritas numa linguagem nova e entendidas por diferentes sistemas. (DZIEKANIAK; KIRINUS, 2004).

3.13.1 W3C

O W3C, é um consórcio mundial, liderado por Tim Berners-Lee, o qual reúne empresas, profissionais, instituições acadêmicas, e cientistas com o intuito de padronizar novas tecnologias que possibilitem estender gradativamente as funcionalidade do ambiente web, levando a internet ao seu potencial máximo.(RAMALHO; VIDOTTI; FUJITA, 2007).

Segundo Souza (2005), o W3C pretende:

> (...)embutir inteligência e contexto nos códigos XML utilizados pra confecção de páginas web, de modo a melhorar a forma com que os programas possam interagir com essas páginas e também possibilitar seu uso mais intuitivo por parte dos usuários.

Com isso, Tim Berners-Lee, pretende a criação de padrões tecnológicos que permitam que programas e dispositivos especializados, chamados agentes, possam interagir entre si, trocando informações, e automatizando as tarefas rotineiras dos usuários. (SOUZA, 2005).

O W3C visa levar a web ao seu potencial máximo através do desenvolvimento de protocolos e recomendações, promovendo a sua evolução e assegurando a interoperabilidade. O W3C tem publicado padrões e diretrizes utilizados para a criação e interpretação de conteúdos na Web. (PANSANATO, 2007)

A web 3.0 seria responsável por definir o significado das palavras e facilitar que um conteúdo na web possua um significado adicional que vá além do significado textual do conteúdo. Desse modo, o W3C define a web 3.0, como uma web extendida, que possui maior significado, onde qualquer usuário pode encontrar respostas de forma mais rápida, através de uma informação melhor definida no momento da busca. (KUSTER; HERNÁNDEZ, 2013).

A partir disso, pretende-se não apenas facilitar as trocas de informações entre os agentes pessoais, mas também estabelecer uma língua comum que possibilite o compartilhamento de dados entre os dispositivos e sistemas de informação. (SOUZA, 2005).

Para atingir esse objetivo, faz-se necessária a padronização de tecnologias, linguagens e metadados, de modo que todos os usuários da web possam obedecer as regras determinadas sobre como armazenar dados e descrever a informação armazenada, fazendo com que ela possa ser compreendida por outros usuários, sejam humanos ou não, de forma automática e sem ambiguidades. (SOUZA, 2005)

O sítio do W3C apresenta alguns princípios listados a seguir:

- Web para todos: visa tornar disponível todos os benefícios da web independetemente do hardware que utilizam, software, infra-estrutura de rede, idioma, cultura, localização geográfica ou capacidade física e mental.
- Web em todas as coisas: permitir o acesso através dos diferentes dispositivos encontrados atualmente.

A visão do projeto W3C, pressupõe: “a participação e o compartilhamento de conhecimentos para gerar confiança em uma escala global.” (w3c.br) Com isso, o W3C pretende ajudar no desenvolvimento das tecnologias que darão suporte a web semântica, viabilizando pesquisas.

Desse modo, na área da web semântica, “várias iniciativas surgiram com o intuito de atualizar os padrões de tesauros internacionais para que considerem esses desenvolvimentos online” (RAMALHO; VIDOTTI; FUJITA, 2007). Junto a isso, o W3C, tem participado do desenvolvimento de padrões que dêem suporte ao uso de *Knowlegde Organization Systems* (Sistemas de Organização do Conhecimento).

Esse software representa desde estruturas simples até estruturas mais complexas e abrangentes, gerando um modelo capaz de expressar a estrutura básica e o conteúdo de tesauros, lista de cabeçalhos, esquemas de classificação entre outros esquemas conceituais. (RAMALHO; VIDOTTI; FUJITA, 2007).

3.13.2 Padrões da web semântica

Para melhor compreendê-la, a seguir são apresentados os padrões e tecnologias adotadas pelo W3C.

- RDF: formato de representação de metadados; trata-se de um dos mais importantes padrões, é a primeira linguagem de representação do conhecimento na web semântica. Trabalha com um trio de informação, o qual expressa o seu significado. Cada componente do trio tem sua própria finalidade, em analogia ao sujeito, verbo e objeto de uma frase e recebe uma identificação URI. Pode ser processado de diversas maneiras por máquinas, tornando-se bases de conhecimento. (BAX, 2013)
- URI: como dito anteriormente, consiste em um identificador único de recursos, que permite a definição e adoção de maneira precisa, de nomes aos recursos e seus respectivos endereços na internet;
- UNICODE: esquema padronizado de codificação de caracteres, diminui consideravelmente a possibilidade de redundância dos dados.
- Signature: tecnologias desenvolvidas para substituir em ambiente computacional a função exercida pela assinatura formal de uma pessoa em suporte físico.
- Encryption: processo em que as informações são criadas de modo que não possam ser interpretadas por qualquer pessoa ou sistema computacional;
- XML : adota o XML, recomendado formalmente pelo W3C, deriva do padrão SGML, e contem *tags* para descrever o conteúdo do documento, seu foco está na descrição dos dados que o documento

contém; é um padrão flexível, podendo-se acrescentar novas *tags* conforme seja necessário. (SOUZA, 2005)

- OWL: linguagem computacional para o desenvolvimento de ontologias
- Nomespace: coleção de nomes, utilizados em documentos XML para validar elementos e atributos;
- Dublin core, que é uma iniciativa para criação de um conjunto de metadados para descrição de documentos eletrônicos, possui 15 elementos e se baseia no padrão MARC. (SOUZA, 2005)
- Trust: camada de confiança, onde se espera garantir que as informações estejam representadas de modo correto, possibilitanto um maior grau de confiabilidade.

O uso de ontologias se dá por meio do OWL (Web Ontology Language). São explicitadas em um documento, e "definindo formalmente as relações entre termos e conceitos, e também as relações entre conceitos em si." (SOUZA, 2005). A ontologia fornece suporte para a evolução de vocabulários e para o processamento e a integração da informação existente sem problemas de indefinição ou conflito de terminologia. (PANSANATO, 2007)

De acordo com o Semanticweb.org (2003, apud Souza, 2005):

> " uma ontologia é uma especificação de uma conceituação. É designada com o propósito de habilitar o compartilhamento e reuso de conhecimentos, de forma a criar ' compromissos ontológicos', ou definições necessárias à criação de um vocabulário comum".

Apesar de ter sua origem na área da filosofia, segundo Guarino (1998, *apud* ROCHA, 2004), no contexto de inteligência artificial a ontologia é definida como uma especificação explícita e até mesmo formal de uma conceitualização compartilhada. Essa conceitualização, segundo o mesmo autor, é uma visão abstrata e sistemática do mundo que se pretende representar.

Conforme Lima-Marques (2006, *apud* NARUKAWA, 2011), na área de inteligência artificial, a ontologia surge como uma possibilidade de compartilhamento e reutilização de conhecimento representado formalmente para uso em sistemas de computador, exigindo a definição de um vocabulário comum para representar este conhecimento.

O objetivo principal da construção de uma ontologia, é possibilitar a troca de informações entre os membros de uma comunidade, seja ela de agentes

humanos ou máquinas inteligentes. Para isso faz-se necessário o uso de terminologias compartilhadas e uma definição de entidades e relacionamentos. (SOUZA, 2005)

Uma ontologia define conceitos e as relações semânticas que se obtém entre esses conceitos, de modo que ela fornece suporte para o processamento de recursos baseado na interpretação do significado do conteúdo. (FERREIRA, 2006).

Segundo Ferreira (2006), as ontologias serão articuladas entre si através de ferramentas denominadas meta-ontologias. Com uma definição padronizada de indexação, é possível criar sistemas que sejam mais funcionais para a recuperação da informação nos ambientes digitais.

Ainda de acordo com a mesma autora, a representação e indexação de um documento deve ser específica o suficiente para explicitar a semântica do seu conteúdo, fazendo uso de tecnologias e padrões que proporcionem automação, compartilhamento, integração e reuso de informação

As ontologias podem ser aplicadas nos sistemas de indexação automática dando suporte para a organização, representação e recuperação da informação nos sistemas, favorecendo a contextualização de informações. (NARUKAWA, 2011).

Desse modo, a web semântica consiste em uma evolução no modo de organização das informações no ambiente web, possibilitando a inclusão de aspectos semânticos aos dados, proporcionando uma melhor busca e recuperação da informação em meio digital para os usuários da web.

4. Resultados

São apresentados os resultados obtidos a partir da análise das 51 publicações para a construção do trabalho, pertinentes a área de indexação automática. As 51 publicações mencionadas são a base deste trabalho e foram retiradas de diferentes bases de dados nacionais e estrangeiras, abarcando o período de 1981 a 2013. Foram efetuadas comparações com a tese de doutorado de Ladeira (2010), onde a autora realizou um estudo sobre a área de PLN dos anos de 1973 a 2009, na base de dados ARIST abarcando apenas publicações nacionais.

4.1 Evolução da Indexação Automática no tempo

A indexação automática evoluiu consideravelmente ao longo do tempo. Para compreendê-la é preciso saber as origens da indexação de modo geral. Segundo Silva e Fujita(2004) foi a partir da atividade de elaboração de índices. Kobashi (1994) afirma que a indexação aparece no século XVII, com a edição de um jornal chamado ‘ *Le Journal des Sçavans*’ publicado no ano de 1667.

A indexação em grande escala surgiu com a compilação da primeira concordância completa da Bíblia por Alexandre Cruden em 1737. (COLLISON, 1971, apud SILVA, FUJITA, 2004). Contudo, foi no século XIX que ela passou a ser vista como uma atividade necessária em meio ao aumento da massa documental. E nesse mesmo período, o tratamento dos índices evoluiu de forma significativa.

Em meio a tantas definições apresentadas, a indexação pode ser entendida, de modo geral, como a atividade que seleciona os termos que melhor representam o conteúdo de um documento, tendo a finalidade de auxiliar o usuário na busca e recuperação da informação que precisa nos diversos sistemas de informação.

O trabalho foca seu estudo na indexação automática, sendo esta compreendida como a atividade que não necessita de uma intervenção direta por parte do documentalista/bibliotecário. Desse modo, a indexação automática surge a partir da necessidade de substituição de um indexador humano por um software.

Seu início se dá em 1948, quando George Zipf formulou duas leis sobre a distribuição das palavras em um texto. A primeira lei se referia a palavras de alta frequência, e a segunda lei as palavras de baixa frequência. Com isso, sua segunda lei foi aperfeiçoada por Booth, ficando conhecida comumente como lei de Zipf-Booth.

Em seguida, por volta da década de 50, Hans Peter Luhn, surge como iniciante na área de estudos bibliométricos relacionados com a frequência de ocorrência das palavras, propondo que a frequência das palavras em um documento está relacionada com sua utilidade no processo de indexação. Em 1958, Baxandale também contribui para a área, comparando a eficiência de três métodos automáticos de indexação.

Em meados de 1959, surge o índice KWIC, onde cada palavra-chave que aparece no título do documento, se torna uma entrada do índice. Faz uso de uma lista de palavras vazias, e é um método que não usa tesauro ou dicionário. O índice KWOC surge de maneira semelhante ao KWIC, contudo as palavras-chaves escolhidas como ponto de acesso são repetidas fora do contexto. A repeito do KWAC, ele funciona da mesma maneira que o KWOC, porém a palavra destacada não é substituída por um sinal gráfico.

A partir da década de 70, tem-se uma intensificação de pesquisas na área de indexação automática, surgindo métodos inteiramente automáticos, como o SMART, onde são introduzidos no sistema computacional trechos do documento, e a partir daí vários procedimentos automáticos de análise do texto são realizados. Outro método é o MEDlars, que faz uso de vocabulário controlado e uma lista de palavras-chaves determinadas com termos de busca e formulações booleanas.

Desse modo, a indexação automática continua como uma importante área de estudo, uma vez que seus problemas de sintaxe e semântica ainda não puderam ser completamente resolvidos. Como alternativa, surge por volta da década de 60 estudos na área de aplicação do processamento de linguagem natural na indexação automática. (NEVES, 2009)

Por meio da sintaxe e da semântica é possível que o software identifique a estrutura lexical e gramatical nas frases e também o significado dos termos que estão representando o conteúdo do documento. Contudo, a manipulação dessas áreas da linguística permance como um desafio para as futuras

gerações, pois ainda não foi possível o seu controle da forma idealizada pelos autores.

4.2 Crescimento da área

De acordo com os dados coletados e a análise realizada, observou-se que a maior parte dos artigos encontrados foram nacionais, mostrando um crescimento significativo de estudos da área no Brasil. Contudo, a seção de experimentos em processamento de linguagem natural, apresentou um número bem maior de experimentos estrangeiros.

Em um total de 20 experimentos apresentados, 15 foram estrangeiros e apenas 5 foram nacionais. Isso indica que apesar de existir uma preocupação com o tratamento e recuperação da informação por parte dos profissionais de diversas áreas, ainda são necessários maiores estudos experimentais no país.

4.3 Projetos e experimentos

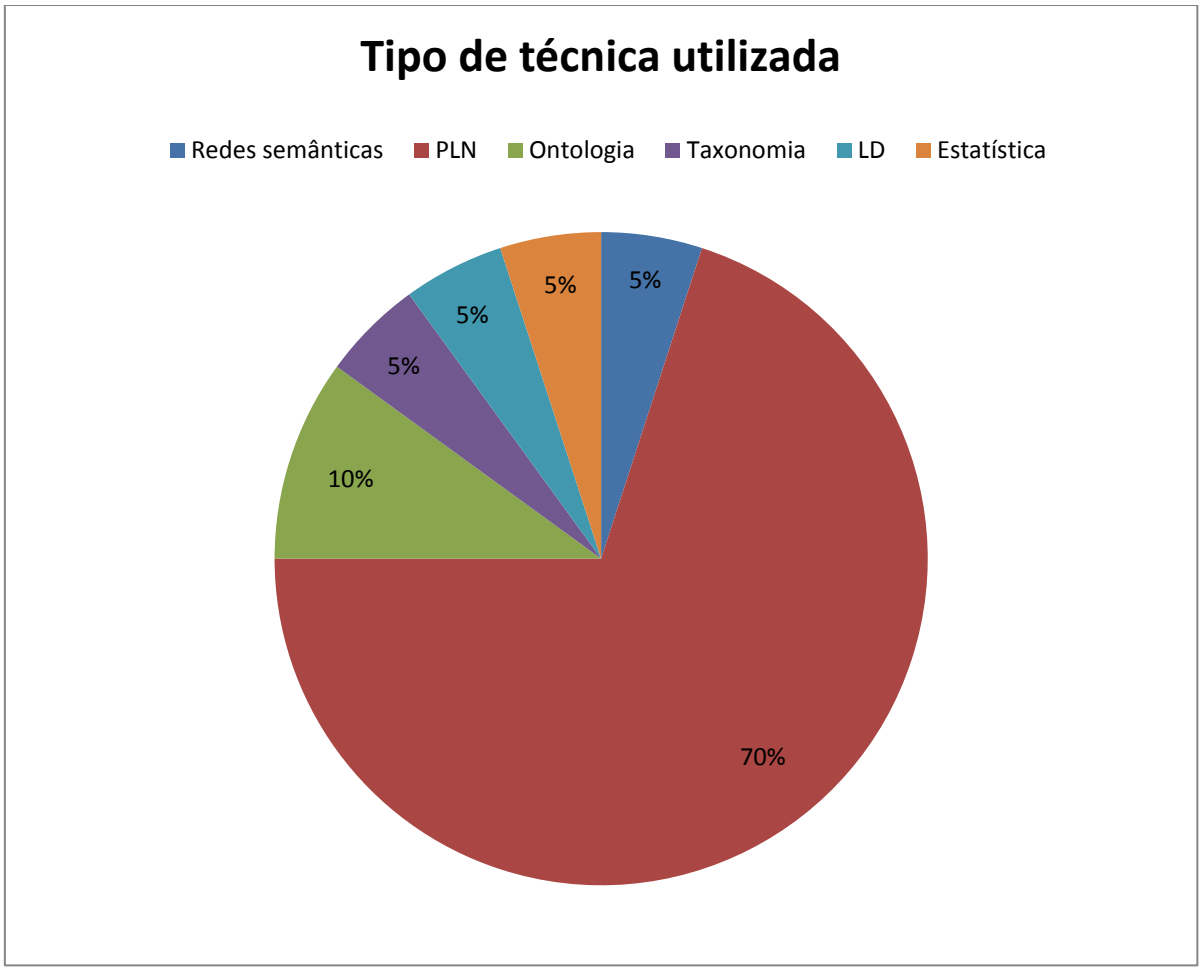
Foi apresentada na sessão 3.9 uma tabela com os experimentos de processamento em linguagem natural, em seguida foi dada um explicação sintética de cada experimento e seu funcionamento. A seguir está uma tabela quantificando os tipos de técnicas utilizadas pelos experimentos.

Tabela 4 – Principais técnicas reveladas a partir da análise de conteúdo

Técnica	Quantidade
Estatística	1
Linguagem documentária	1
Ontologia	2
Processamento de linguagem natural	14
Redes semânticas	1
Taxonomia	1

Fonte: elaboração própria

Gráfico 1 – Tipo de técnica utilizada



Fonte: elaboração própria

A partir dos dados apresentados, é possível notar que a técnica referente ao processamento de linguagem natural foi a mais utilizada na maior parte dos experimentos apresentados, essa técnica aborda também os conceitos de sintaxe e semântica. Segundo apresentado por Ladeira (2010), a autora encontrou as seguintes técnicas relacionadas com três áreas distintas do conhecimento:

1. Ciência da computação: gramática, *parser*, *corpus*;
2. Processamento de Linguagem Natural: léxico, *parser*, *corpus*;
3. Ciência da informação: tesauro.

Importante salientar, que Ladeira (2010) realizou o seu estudo baseando-se em materiais de lingua portuguesa, o que refina consideravelmente o campo de busca.

4.4 Análise comparativa

O presente trabalho abarcou o período de 1981 a 2013, os últimos 32 anos da evolução da indexação automática. Foram realizadas pesquisas em bases de dados e encontrados diversos tipos de documentos, como artigos, dissertações, entre outros.

A partir dos dados da tabela a seguir, realiza-se a divisão quantificada dos documentos por tipo, e também em nacionais e estrangeiros. Desse modo, a partir da tabela 5 é possível notar que a maior parte dos documentos encontrados foram artigos:

Tabela 5 – Tipo de documento e origem

Tipo de documento	Nacionais	Estrangeiros	Total
Artigo	31	8	39
Dissertação	4	0	4
Livro	1	1	2
Tese	6	0	6
Total	36	8	**51**

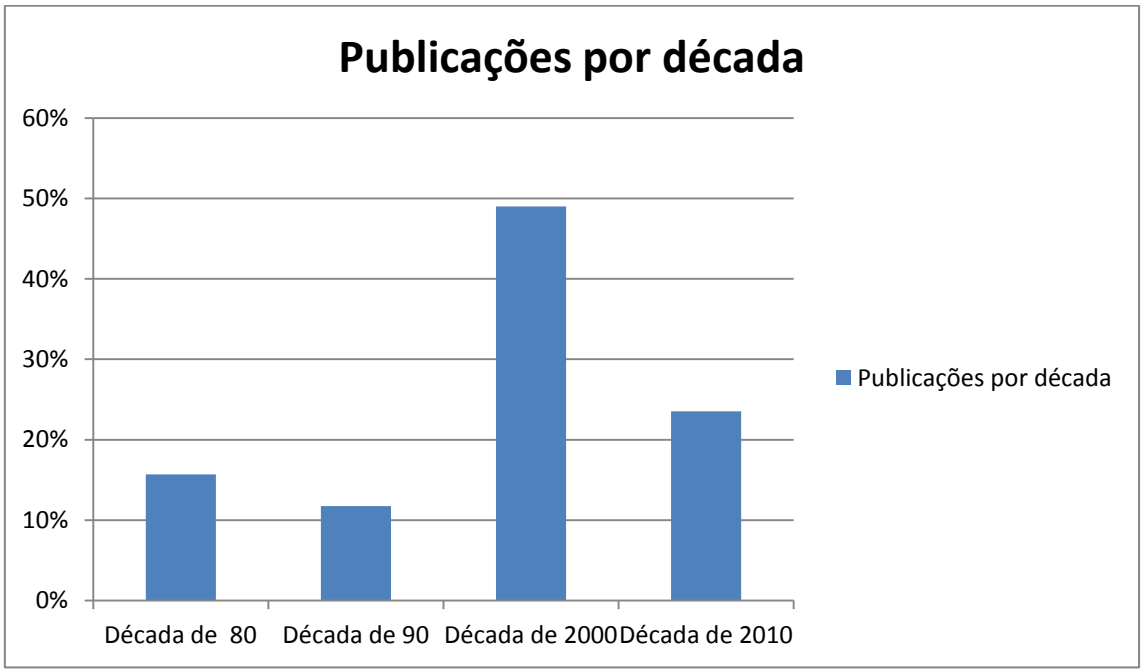
Fonte: elaboração própria

Apesar da busca ter sido realizada também em bases de dados estrangeiras, a maior parte dos documentos são nacionais. As pesquisas foram realizadas utilizando os seguintes termos de busca: ‘ indexação’,‘ indexação automática’, ‘processamento de linguagem natural’, ‘indexação semi-automática’, ‘*automatic indexing*’ e ‘*natural language processing*’ ; o que acabou gerando uma quantidade maior de documentos em língua portuguesa.

No anexo 1, seguindo o modelo de Ladeira (2010), estão listados os documentos utilizados para o presente trabalho, contudo optou-se por organizá-los de acordo com a década em que foram publicados, iniciando na década de 80 até a década de 2010. Nota-se que na década de 2000, foi o período em que ocorreu o maior número de publicações na área, com um total

de 49% dos documentos. Como apresentado em porcentagem no gráfico a seguir:

Gráfico 2- Publicações por década



Fonte: elaboração própria

Fazendo uma análise comparativa com os resultados apresentados por Ladeira (2010) em sua tese de doutorado, foi possível notar que a autora encontrou cerca de 70% dos seus documentos concentrados na década de 2000, contudo vale ressaltar que a autora abarcou o período de 1973 até 2009, não contemplando a década de 2010.

Outra comparação realizada diz respeito as referências apresentadas neste trabalho e as utilizadas por Ladeira (2010) em sua tese. Após uma análise, observou-se que apenas dois documentos em comum foram encontrados nos dois trabalhos. Apresentados a seguir:

- CHOWDHURY, G. Natural language processing. *Annual Review of Information Science and Technology*, n. 37, p. 51-89, 2003.

Trata-se de obra especializada na área de Ciência da Informação.

- LANCASTER, F. W. *Indexação e resumos*: teoria e prática. 2 ed. Brasília: Briquet de Lemos, 2004. 451 p.

  Trata-se de obra clássica quando a temática estudada diz respeito a indexação.

Esse resultado se deve entre outros motivos, ao fato de que Ana Paula focou seu estudo na área de PLN e na base ARIST, já o presente trabalho teve como temática o estudo da indexação automática, e as buscas ocorreram em diversas bases de dados.

Ainda assim, as duas obras mencionadas possuem um valor significativo na área de estudo abarcada por este trabalho. A primeira trata-se de obra especializada na área de Ciência da Informação; a segunda, é um clássico na temática de indexação. Desse modo, os dois conteúdos são essenciais para a construção de trabalhos que abordem tais temas.

5 Conclusão

O trabalho mostrou que a indexação automática teve sua origem a muitos anos, por volta da década de 1950. Foram recuperados artigos desde de 1981 até 2013, mostrando que a área está em uma constante evolução. Contudo, notou-se que o maior volume de trabalhos foram encontrados na década de 2000.

Verificou-se na área de Biblioteconomia e Ciência da Informação os estudos relacionados com indexação automática, obtendo-se um total de 51 publicações encontradas em bases nacionais e estrangeiras. Foram apresentados experimentos realizados na temática de indexação automática e processamento de linguagem natural, obtendo-se um total de 20 experimentos, entre 5 nacionais e 15 estrangeiros.

Optou-se ainda por uma comparação com os resultados obtidos na tese de doutorado de Ladeira (2010), onde a autora focou seu estudo em trabalhos brasileiros e na temática de processamento da linguagem natural. De acordo com o estudo realizado por Ladeira (2010), a autora fez um levantamento de publicações nacionais coletadas automaticamente da Plataforma Lattes, utilizando um instrumento de seleção automática, construído a partir da análise de assunto dos artigos de revisão da base de dados ARIST obtendo uma amostra de 68 trabalhos nacionais publicados no período de 1973 até 2009. Com isso, foi realizada uma análise de conteúdo em cima dos documentos.

Como resultado notou-se que a técnica referente ao processamento de linguagem natural foi a mais encontrada nos experimentos apresentados neste estudo, o que se assemelha as técnicas apresentadas por Ladeira (2010), onde a autora encontrou técnicas de PLN, mas também de outros áreas do conhecimento.

Desse modo, corroborando com os dados apresentados por Ladeira (2010), observou-se que a maior parte dos documentos analisados tanto neste trabalho como também na tese de doutorado apresentado pela autora, foram publicados na década de 2000.

Além disso, por meio da análise de conteúdo, a autora observou ainda que a Ciência da Informação priorizou pesquisas com enfoque na indexação

automática, depois na análise de conteúdo; e a recuperação da informação foi uma problemática de destaque na produção científica nacional.

Foram brevemente mencionados os conceitos de recuperação da informação e web semântica no contexto de indexação automática, mostrando que a indexação interfere diretamente no crescimento dessas temáticas, além de serem assuntos em ascensão na atualidade.

Com isso, o presente trabalho mostrou, através da revisão de literatura, parte do que já foi publicado e o que está sendo foco de estudo na temática de indexação automática, quais são as principais lacunas encontradas e como se deu a evolução da área ao longo dos anos até os dias atuais.

Ladeira (2010) afirma em sua pesquisa que a ciência da informação encontra-se muito tímida em seus estudos, deixando para a ciência da computação e para a linguística o estudo dessa temática. Outro ponto importante para o presente trabalho é o fato de que a autora pôde constatar que a ciência da informação tem dado prioridade para pesquisas relacionadas com a indexação automática, foco deste trabalho.

A situação da indexação automática foi apresentada, procurando selecionar documentos dos principais autores e também aqueles publicados recentemente. Notou-se que a indexação automática é uma área de suma importâcia para a recuperação da informação, e com isso tem voltado a ser foco de diversos estudos que tentam abordar, principalmente, a problemática da semântica, a qual ainda não foi solucionada.

Atualmente, a indexação automática tem voltado a ser foco de diversas pesquisas, pois a recuperação da informação depende diretamente do bom funcionamento da indexação. Procura-se atingir um nível satisfatório na busca realizada pelos usuários, de modo que possam encontrar aquilo que procuram de forma mais fácil e rápida.

Em meio ao contexto da internet, que tem ganhado cada vez mais espaço na busca de informações, a web semântica surge com a tentativa de proporcionar um conteúdo significativo na web, aumentando a eficiência dos motores de busca atuais, contando com a indexação automática para auxiliá-la na escolha dos descritores que irão representar os diferentes tipos de documentos na recuperação da informação.

6 Bibliografia

ALCAIDE, G. S, et. al. Análise comparativa e de consistência entre representações automática e manual de informações documentárias. **Transinformação**, Campinas, v. 13, n. 1, p. 23-41, 2011.

ANDREEWSKI, A., RUAS, V. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ci. Inf**., Brasília, n. 12, p. 61-73, 1983.

BARANOW, U. G. Perspectivas na contribuição da linguística e de áreas afins à Ciência da Informação. **Ci. Inf**., Brasília, n. 12, p. 23-35, 1983.

BARROS, A. J. S; LEHFELD, N. A. S. **Fundamentos de metodologia científica**: um guia para a iniciação científica. 2 ed. São Paulo: Makron Books, 2000, 122 p.

BAX, M. P. A evolução da Web rumo à web semântica. **Prisma.com**, Minas Gerais, n. 19, 2013.

BEHERA, A., LALANNE, D., INGOLD, R. DocMIR: na automatic document-bases indexing system for meeting retrieval. **Bussiness Media**, Fribourg, 2007.

BORGES, G. S. B. ; MACULAN, B. C. M. S. ; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação e Sociedade**, João Pessoa, v. 18, n. 2, p. 181-193, maio/ago. 2008.

BORGES, G. S. B**. Indexação automática de documentos textuais**: proposta de critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais. 2009.

BRASCHER, M. A ambiguidade na recuperação da informação. **Revista de Ciência da Informação**, v. 3, n. 1, 2002.

BRITO, M. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ci. Inf.**, Brasília, n. 21, p. 223-232, 1992.

BRUZINGA, G. S.; MACULAN, B. C. M. S.; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. In: Encontro Nacional de Pesquisa em Ciência da Informação, 8, 2007. Salvador. **Anais...**

CAFÉ, L. BRASCHER, M. Organização do conhecimento: teorias semânticas como base para estudo e representação de conceitos. **Inf. Inf**., Londrina, v. 16, p. 25-51, 2011.

CÂMARA JÚNIOR, A. T. **Indexação automática de acórdãos por meio de processamento de linguagem natural.** 2007. 141 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília. 2007.

CHOWDHURY, G. Natural language processing. **Annual Review of Information Science and Technology**, n. 37, p. 51-89, 2003.

CONTERATTO, G. B. H. Semântica e computação: uma interação necessária**. Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 353-367, 2006.

CÔRREA, R. F., MIRANDA, D. G., LIMA, C. O. A., et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **Novas práticas em informação e conhecimento**, Curitiba, v. 1, n. 1, 2011.

CUNHA, Murilo Bastos da. Desafios na construção de uma biblioteca digital. **Ci. Inf., Brasilia** , v. 28, n. 3, Dec. 1999 .

DA ROCHA, R. P. Metadados, Web Semântica, categorização automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web. **Em Questão**, Porto Alegre, v. 10, n. 1, p. 109-121, 2004.

DUQUE, C. G. SRILiCO: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia. 2005. 118 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais, 2005.

DURAN, N. D., *et al.* The linguistic correlates os conversational deception: comparing natural language processing Technologies. **Applied Psycholinguistics**, n. 31, p. 439-462, 2009.

DZIEKANIAK, G. V., KIRINUS, J. B. Web semântica. **R. Eletr. Bibliotecon. Ci. Inf.**, Florianópolis, n. 18, 2004.

FERREIRA, E. C. H. G. **Geração automática de metadados**: uma contribuição para a Web Semântica. 2006. 228 f. Tese (Doutorado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo. 2006.

GIL LEIVA, I.; RODRIGUEZ MUÑOZ, J. V. El procesacmiento del linguaje natural aplicado al análisis del contenido de los documentos**. Revista General de Información y Documentación**, Madrid, v. 6, n. 2, 1996.

GUEDES, R. M.; MOURA, M. A.; DIAS, E. J. Indexação social e pensamento dialógico: reflexões teóricos. **Inf. Inf**., Londrina, v.16, n. 3, 2011.

GUEDES, V. L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ci. Inf**., Brasília, v. 23, n. 3, p. 318-326, 1994.

HALLER. J. Análise automática de textos em sistemas de informação. **R. Bibliotecon. Brasília,** n. 11, p. 105-113, 1983.

HOLANDA, C. ; BRAZ, M. I. Indexação automática de conteúdos na web: análise de sites de museus. **Biblionline***,* João Pessoa, v. 8, n. 1, p. 42-59, 2012.

KURAMOTO, H. Proposta de um Sistema de Recuperação de Informação assistido por Computador – SRIAC. **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n. 2, p. 211-228, 1997.

KURANZ, J.; GILLES, B. Indexing electronic medical records using a taxonomy. **Bulletin of the American Society for Information Science and Technology**, v. 39, n. 2, 2013.

KUSTER, I.; HERNÁNDEZ, A. **De la web 2.0 a la web 3.0**: antecedentes y consecuencias de la actitud e intención de uso de las redes sociales en la web semántica. Universia Business Review., Valencia, 2013.

LADEIRA, A. P. **Processamento de linguagem natural**: caracterização da produção científica dos pesquisadores brasileiros. 2010. 259 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Minas Gerais, 2005.

LANCASTER, F. W. Indexação e resumos: teoria e prática. 2 ed. Brasília: Briquet de Lemos, 2004. 451 p.

LIMA, V. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em Ciência da informação do vocabulário controlado do SIBI/USP nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, 2009.

MAIA, L. C.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, 2010.

MAMFRIM, F. P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa**. Ci. Inf.** , Brasília, n. 20, p. 191-203, 1991.

NARUKAWA, C. M. Estudo de vocabulário controlado na indexação automática: aplicação no processo de indexação do sistema de idizacion semiautomatica (SISA). 2011. 224 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciência, Universidade Estadual Paulista, São Paulo, 2011.

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**., João Pessoa, v. 19, n. 2, p. 99-118. 2009.

NEVES, Viviane. **Indexação automática de documentos textuais**: iniciativa dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72 f. Tese (Graduação em Biblioteconomia) - Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Arte, Universidade de São Paulo, São Paulo. 2009.

OTHERO, G. A. Linguística Computacional: uma breve introdução. **Letras de hoje**, Porto Alegre, v. 41, n. 2, p. 341-351, 2006.

PANSANATO, L. T. E. **Um modelo de navegação exploratória para a infra-estrutura da Web Semântica**. 2007. 194 f. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, USP, São Paulo, 2007.

PÉREZ, D., et. al. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. **Revista Signos**, Madrid, n. 38, p. 325-343, 2005.

RAMALHO, R. A. S., VIDOTTI, S. A. B. G., FUJITA, M. S. L. Web semântica: uma investigação sob o olhar da Ciência da Informação. **Revista de Ciência da Informação**, v. 8, n. 6, 2007.

ROBREDO, J. **A indexação automática de textos**: o presente já entrou no futuro. In: JAIME ROBREDO. Brasília: Universidade de Brasília, p. 235-274.

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. Brasília: edição de autor, 2005, 409 p.

SALINAS ORDOÑEZ, S., GELBUKH, A. Representación computacional del lenguaje natural escrito. **Ingeniería**, v. 15, n. 1, p. 6-21, 2010.

SAUTCHUK, I. **Prática da morfossintaxe**: como e por que aprender análise (morfo)sintática. 2 ed. Barueri: Manole, 2010, 250 p.

SILVA, M. R. ; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, p. 133-161, maio/ago.,2004.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 215 f. Tese (Doutorado em Ciência da informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais, 2005.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. In: Encontros Bibli, 1, 2006. Florianópolis. **Anais....**Universidade Federal de Santa Catarina, 2006.

TAMBA-MECZ, I. **A semântica**. São Paulo: Parábola Editorial, 2006, 159 p.

UNISIST. Princípios de indexação. **R. Esc. Bibliotecon.**, Belo Horizonte, n. 10, p. 83-94, 1981.

VIEIRA, S. B. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **Revista de Biblioteconomia, Brasília**, v. 16, p. 83-94, jan./jun. 1988.

VIEIRA, S. B. **Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação**. 1984. 204 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília, 1984.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ci Inf.** , Brasília, n. 17, p. 43-57, 1988.

7 Apêndice

Apêndice 1

Lista cronológica, separada por década do material coletado durante o trabalho.

Década de 1980

1. UNISIST. Princípios de indexação. **R. Esc. Bibliotecon.**, Belo Horizonte, n. 10, p. 83-94, 1981.
2. ANDREEWSKI, A., RUAS, V. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ci. Inf**., Brasília, n. 12, p. 61-73, 1983.
3. BARANOW, U. G. Perspectivas na contribuição da linguística e de áreas afins à Ciência da Informação. **Ci. Inf**., Brasília, n. 12, p. 23-35, 1983.
4. HALLER. J. Análise automática de textos em sistemas de informação. **R. Bibliotecon. Brasília,** n. 11, p. 105-113, 1983.
5. ROBREDO, J. A indexação automática de textos: o presente já entrou no futuro. In: JAIME ROBREDO. Brasília: Universidade de Brasília, p. 235-274. [198?]
6. VIEIRA, S. B. **Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação**. 1984. 204 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília, 1984.
7. VIEIRA, S. B. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **Revista de Biblioteconomia, Brasília**, v. 16, p. 83-94, jan./jun. 1988.
8. VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ci Inf.** , Brasília, n. 17, p. 43-57, 1988.

Década de 1990

1. MAMFRIM, F. P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa**. Ci. Inf.** , Brasília, n. 20, p. 191-203, 1991.
2. BRITO, M. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ci. Inf.**, Brasília, n. 21, p. 223-232, 1992.
3. GUEDES, V. L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ci. Inf**., Brasília, v. 23, n. 3, p. 318-326, 1994.
4. GIL LEIVA, I. RODRIGUEZ MUÑOZ, J. V. El procesacmiento del linguaje natural aplicado al análisis del contenido de los documentos**.**

Revista General de Información y Documentación, Madrid, v. 6, n. 2, 1996.
5. KURAMOTO, H. Proposta de um Sistema de Recuperação de Informação assistido por Computador – SRIAC. **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n. 2, p. 211-228, 1997.
6. CUNHA, Murilo Bastos da. Desafios na construção de uma biblioteca digital. **Ci. Inf., Brasilia** , v. 28, n. 3, Dec. 1999.

Década de 2000

1. BRASCHER, M. A ambiguidade na recuperação da informação. **Revista de Ciência da Informação**, v. 3, n. 1, 2002.
2. CHOWDHURY, G. Natural language processing. **Annual Review of Information Science and Technology**, n. 37, p. 51-89, 2003.
3. DZIEKANIAK, G. V., KIRINUS, J. B. Web semântica. **R. Eletr. Bibliotecon. Ci. Inf.**, Florianópolis, n. 18, 2004.
4. ROCHA, R. P. Metadados, Web Semântica, categorização automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web. **Em Questão**, Porto Alegre, v. 10, n. 1, p. 109-121, 2004.
5. SILVA, M. R. ; FUJITA, M. S. L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, p. 133-161, maio/ago.,2004.
6. LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2 ed. Brasília: Briquet de Lemos, 2004. 451 p.
7. DUQUE, C. G. SRILiCO: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia. 2005. 118 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais, 2005.
8. 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. Brasília: edição de autor, 2005, 409 p.
9. PÉREZ, D., *et. al*. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. **Revista Signos**, Madrid, n. 38, p. 325-343, 2005.
10. CONTERATTO, G. B. H. Semântica e computação: uma interação necessária**. Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 353-367, 2006.
11. FERREIRA, E. C. H. G. **Geração automática de metadados**: uma contribuição para a Web Semântica. 2006. 228 f. Tese (Doutorado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo. 2006.

12. SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. In: Encontros Bibli, 1, 2006. Florianópolis. **Anais....**Universidade Federal de Santa Catarina, 2006.
13. SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. 2005. 215 f. Tese (Doutorado em Ciência da informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais, 2005.
14. OTHERO, G. A. Linguística Computacional: uma breve introdução. **Letras de hoje**, Porto Alegre, v. 41, n. 2, p. 341-351, 2006.
15. BEHERA, A., LALANNE, D., INGOLD, R. DocMIR: na automatic document-bases indexing system for meeting retrieval. **Bussiness Media**, Fribourg, 2007.
16. BRUZINGA, G. S.; MACULAN, B. C. M. S.; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. In: Encontro Nacional de Pesquisa em Ciência da Informação, 8, 2007. Salvador. **Anais...**
17. CÂMARA JÚNIOR, A. T. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. 141 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília, Brasília. 2007.
18. RAMALHO, R. A. S.; VIDOTTI, S. A. B. G.; FUJITA, M. S. L. Web semântica: uma investigação sob o olhar da Ciência da Informação. **Revista de Ciência da Informação**, v. 8, n. 6, 2007.
19. PANSANATO, L. T. E. Um modelo de navegação exploratória para a infra-estrutura da Web Semântica. 2007. 194 f. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, USP, São Paulo, 2007.

20. BORGES, G. S. B. ; MACULAN, B. C. M. S. ; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação e Sociedade**, João Pessoa, v. 18, n. 2, p. 181-193, maio/ago. 2008.
21. BORGES, G. S. B. Indexação automática de documentos textuais: proposta de critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Minas Gerais. 2009.
22. DURAN, N. D., et al. The linguistic correlates os conversational deception: comparing natural language processing Technologies. **Applied Psycholinguistics**, n. 31, p. 439-462, 2009.
23. LIMA, V. M. A., BOCCATO, V. R. C. O desempenho terminológico dos descritores em Ciência da informação do vocabulário controlado do SIBI/USP nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, 2009.

24. 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**., João Pessoa, v. 19, n. 2, p. 99-118. 2009.
25. NEVES, Viviane. **Indexação automática de documentos textuais**: iniciativa dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72 f. Tese (Graduação em Biblioteconomia) - Departamento de Biblioteconomia e Documentação da Escola de Comunicações e Arte, Universidade de São Paulo, São Paulo. 2009.

Década de 2010

1. MAIA, L. C., SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, 2010.
2. LADEIRA, A. P. Processamento de linguagem natural: caracterização da produção científica dos pesquisadores brasileiros. 2010. 259 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Minas Gerais, 2010.
3. ALCAIDE, G. S., et. al. Análise comparativa e de consistência entre representações automática e manual de informações documentárias. **Transinformação**, Campinas, v. 13, n. 1, p. 23-41, 2011.
4. BAX, M. P. A evolução da Web rumo à web semântica. **Prisma.com**, Minas Gerais, n. 19, 2013.
5. CAFÉ, L.; BRASCHER, M. Organização do conhecimento: teorias semânticas como base para estudo e representação de conceitos. **Inf. Inf**., Londrina, v. 16, p. 25-51, 2011.
6. CÔRREA, R. F., et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **Novas práticas em informação e conhecimento**, Curitiba, v. 1, n. 1, 2011.
7. GUEDES, R. M.; MOURA, M. A.; DIAS, E. J. Indexação social e pensamento dialógico: reflexões teóricos. **Inf. Inf**., Londrina, v.16, n. 3, 2011.
8. HOLANDA, C. ; BRAZ, M. I. Indexação automática de conteúdos na web: análise de sites de museus. **Biblionline***,* João Pessoa, v. 8, n. 1, p. 42-59, 2012.
9. KURANZ, J.; GILLES, B. Indexing electronic medical records using a taxonomy. Bulletin of the American Society for Information Science and Technology, v. 39, n. 2, 2013.
10. KUSTER, I.; HERNÁNDEZ, A. **De la web 2.0 a la web 3.0**: antecedentes y consecuencias de la actitud e intención de uso de las redes sociales en la web semántica. Universia Business Review., Valencia, 2013.
11. NARUKAWA, C. M. **Estudo de vocabulário controlado na indexação automática**: aplicação no processo de indexação do sistema de

idizacion semiautomatica (SISA). 2011. 224 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciência, Universidade Estadual Paulista, São Paulo, 2011.
12. SALINAS ORDOÑEZ, S.; GELBUKH, A. Representación computacional del lenguaje natural escrito. **Ingeniería**, v. 15, n. 1, p. 6-21, 2010.