UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE ARTES E COMUNICAÇÃO
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
MESTRADO EM CIÊNCIA DA INFORMAÇÃO

Victor Galvão Celerino

PROPOSTA DE NORMALIZAÇÃO DOS SINTAGMAS NOMINAIS EM TERMOS PARA INDEXAÇÃO AUTOMÁTICA

Recife – PE
2018

VICTOR GALVÃO CELERINO

PROPOSTA DE NORMALIZAÇÃO DOS SINTAGMAS NOMINAIS EM TERMOS PARA INDEXAÇÃO AUTOMÁTICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Ciência da Informação.

Área de concentração: Informação, Memória e Tecnologia.

Linha de pesquisa: Comunicação e Visualização da Memória.

Orientador: Prof. Dr. Renato Fernandes Corrêa

Recife- PE

2018

Catalogação na fonte
Bibliotecário Jonas Lucas Vieira, CRB4-1204

C392p Celerino, Victor Galvão

Proposta de normalização dos sintagmas nominais em termos para indexação automática / Victor Galvão Celerino. – Recife, 2018.

171 f.: il., fig.

Orientador: Renato Fernandes Corrêa.

Dissertação (Mestrado) – Universidade Federal de Pernambuco, Centro de Artes e Comunicação. Ciência da Informação, 2018.

Inclui referências e apêndices.

1. Indexação automática. 2. Sintagmas nominais. 3. Recuperação de informação. 4. Normalização de sintagmas nominais. I. Corrêa, Renato Fernandes (Orientador). II. Título.

020 CDD (22. ed.) UFPE (CAC 2018-68)

Serviço Público Federal
Universidade Federal de Pernambuco
Programa de Pós-graduação em Ciência da Informação - PPGCI

VICTOR GALVÃO CELERINO

Proposta de normalização dos sintagmas nominais em termos para indexação automática

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de mestre em Ciência da Informação.

Aprovada em: 26/02/2018

BANCA EXAMINADORA

_____

Prof. Dr. Renato Fernandes Corrêa (Orientador)
Universidade Federal de Pernambuco

_____

Prof. Dr. Fabio Assis Pinho (Examinador Interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. André Anderson Cavalcante Felipe (Examinador Externo)
Universidade Federal de Pernambuco

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer a Deus por tudo que proporcionou até agora e o que irá proporcionar futuramente.

Em especial, gostaria de agradecer a toda a minha família. Agradeço ao meu Pai e minha Mãe que sempre me apoiaram em todas as decisões que tomei e em todos os momentos difíceis que passei. Agradeço também aos meus irmãos que assim como meus pais também me auxiliaram nessa pesquisa.

Em seguida gostaria de agradecer ao Programa de Pós-graduação em Ciência da Informação da Universidade Federal de Pernambuco, pela grande oportunidade que me foi dada nesses dois anos de pesquisa.

Sou grato também pela oportunidade de ter como orientador o Prof. Dr. Renato Fernandes Corrêa, que com paciência me orientou no desenvolvimento dessa pesquisa nesses dois anos.

Agradeço também a todos os professores que fazem parte do Programa de Pós-graduação em Ciência da Informação da UFPE, pois todos contribuíram para o desenvolvimento dessa pesquisa.

Por fim, sou grato a todos eles pois sem eles essa pesquisa não seria possível.

RESUMO

Atualmente vivenciamos um crescimento informacional extraordinário, entretanto esse crescimento é acompanhado de um grande problema: como organizar toda essa informação? No cenário da organização e recuperação da informação digital, tem se destacado a Indexação Automática através do uso dos Sintagmas Nominais. Diferente da indexação praticada através de palavras isoladas, os Sintagmas Nominais são unidades sintáticas que possuem semântica, ou seja, possuem um sentido específico. Nesse contexto, o presente trabalho tem como objetivo geral propor um método de normalização dos Sintagmas Nominais, extraídos automaticamente, em termos canônicos, para que sejam satisfatórios como descritores dos documentos. No contexto da indexação automática por sintagmas nominais, pesquisas indicaram que nem todos os Sintagmas Nominais podem ser considerados descritores. Portanto, esta pesquisa: investigou os processos ligados a indexação automática por Sintagmas Nominais; selecionou manualmente Sintagmas Nominais contendo palavras-chaves; minimizou Sintagmas Nominais extensos; alterou os Sintagmas Nominais para aproximar de termos descritores; comparou os Sintagmas Nominais normalizados com termos do TBCI; e avaliou o método proposto em um experimento de normalização dos Sintagmas Nominais. Para atingir os objetivos propostos para esta pesquisa, foram utilizados a pesquisa bibliográfica e pesquisa empírica, com a realização da proposição e avaliação de método de normalização por meio da aplicação de um experimento. Através da pesquisa bibliográfica foi possível identificar estudos realizados sobre a indexação automática através de Sintagmas Nominais, estudos esses que auxiliaram no desenvolvimento da proposta de método de normalização dos Sintagmas Nominais. O experimento foi composto de duas etapas. A primeira etapa possui 85 regras voltadas a minimizar os Sintagmas Nominais extensos, e a segunda etapa lida com critérios voltados a alterar a estrutura dos Sintagmas Nominais, aproximando-os de termos canônicos. Através desse experimento foi possível avaliar e determinar quais critérios seriam importantes para a normalização dos Sintagmas Nominais. Os resultados apresentados no experimento indicaram que as etapas 1 e 2 da proposta de normalização foram satisfatórias. Concluiu-se que a proposta de normalização conseguiu atingir seu objetivo, pois os Sintagmas Nominais foram normalizados preservando a sua estrutura e as palavras-chave.

Palavras-chave: Indexação Automática. Sintagmas Nominais. Recuperação de Informação. Normalização de Sintagmas Nominais.

ABSTRACT

Today we are experiencing extraordinary informational growth, but this growth generates a major problem: how can we organize all this information? In the scenario of the organization and retrieval of digital information, Automatic Indexing with using of the Noun Phrases highlights. Unlike the indexation practiced with isolated words, the Noun Phrases are syntactic units that have semantics, that is, they have a specific meaning. In this context, the present work has as general objective to propose a method of normalization of the Noun Phrases extracted automatically in canonical terms, so that they are satisfactory as descriptors of the documents. In the context of automatic indexing by Noun phrases, researches indicate that not all Noun Phrases are descriptors. Therefore, this research: investigated the processes of automatic indexing by Noun Phrases; manually selected Noun Phrases containing the keywords; changed Noun Phrases to approximate descriptor terms; minimized extensive Noun Phrases; compared the normalized Noun Phrases with the TBCI terms; and evaluated the proposed method in a Noun Phrases normalization experiment. In order to reach the objectives proposed for this research, were used bibliographic research and empirical research, with the accomplishment of the proposition and evaluation of method of normalization through the application of an experiment. Through the bibliographic research, it was possible to identify studies carried out on automatic indexing through Noun Phrases, which have helped to develop the proposal of normalization of Noun Phrases. The experiment was composed of two steps. The first step has 85 rules aimed at minimizing the extended Noun Phrases, and the second step deals with criteria aimed at changing the structure of the Noun Phrases, bringing them closer to canonical terms. Through the experiment, it was possible to evaluate and determine which criteria would be important for the normalization of the Noun Phrases. The results presented in the experiment indicated that steps 1 and 2 of the normalization proposal were satisfactory. It concludes that the normalization proposal succeeded in achieving its objective, since the Noun Phrases normalized preserving their structure and the keywords.

Keywords: Automatic Indexing. Noun Phrases. Information Retrieval. Normalization of Noun Phrases.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

CI	Ciência da Informação
OI	Organização da Informação
RI	Recuperação da Informação
SRI	Sistemas de Recuperação da Informação
SN	Sintagma Nominal
SNs	Sintagmas Nominais
RTI	Representação Temática da Informação
OC	Organização do Conhecimento
RC	Representação do Conhecimento
TTI	Tratamento Temático da Informação
LC	Library of Congress
UNISIST	Sistema Mundial de Informação Científica
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
ABNT	Associação Brasileira de Normas Técnicas
LN	Linguagem Natural
LD	Linguagem Documentária
KWIC	Keyword in Context
IBBD	Instituto Brasileiro de Bibliografia e Documentação
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
SV	Sintagmas Verbais
PDF	Portable Document Format
HTML	HyperText Markup Language
SNa	Sintagma Nominal Antes
SNd	Sintagma Nominal Depois

SUMÁRIO

1 INTRODUÇÃO

O crescimento na quantidade de informação disponível atualmente se deve ao advento da escrita, da imprensa e recentemente, aos avanços tecnológicos na eletrônica, informática, internet e comunicações.

A informação é necessária para produzir conhecimento, mas para que isso ocorra é preciso que o usuário possa acessá-la, pois não é suficiente a informação estar disponível em bases de dados e bibliotecas se o usuário não consegue recuperá-la.

Devido a esse crescimento na produção informacional do mundo, ficou em destaque um problema: como identificar, tratar e disponibilizar toda essa informação para os usuários? Nesse momento é que surgem pesquisas com o intuito de tornar essas informações acessíveis ao usuário (MAIA; SOUZA, 2010).

Esse crescimento informacional é apresentado na literatura como sendo o fenômeno da “explosão informacional”. Na década de 1950, surge uma ciência com o objetivo de estudar a informação para assim desenvolver teorias e práticas que solucionassem os problemas ocasionados pela crescente produção de informação, que acontecia de maneira desordenada. Essa ciência ficou conhecida como Ciência da Informação (CI).

Segundo Oliveira (2011), a CI nasceu após uma revolução científica e técnica que ocorreu no período da Segunda Guerra Mundial. Nessa época, diversos países realizaram estudos e desenvolveram tecnologias, em sua maioria voltados a guerra, gerando assim um crescimento no fluxo informacional. Durante esse período histórico foi apresentado o artigo desenvolvido por Vannevar Bush em 1945 intitulado, “*As we may think*”, onde Bush destacou os obstáculos que existiam para a organização e fluxo das informações sigilosas geradas durante a Segunda Guerra Mundial. (SARACEVIC, 1996; BARRETO, 2002).

Atualmente a CI é considerada uma ciência voltada ao estudo da informação, abrangendo suas propriedades, seu comportamento, seu fluxo e seus meios de processamento, com o objetivo de proporcionar a acessibilidade e utilização da informação de forma eficaz (BORKO, 1968, *apud* OLIVEIRA, 2011).

É necessário salientar que a preocupação com a informação, no âmbito do seu armazenamento, organização e recuperação, antecede a CI. Essa preocupação já

advém de muito tempo atrás, na idade média, onde mesmo com um volume de informações bem inferior, se comparado com o cenário atual, já existiam estudiosos que se preocupavam em como lidar com a informação e desenvolviam métodos e técnicas voltados a guarda e organização. Entretanto, segundo Saracevic (1996), a CI vai além do simples tratamento e recuperação da informação, pois ela se preocupa com os processos de comunicação humana.

> [...] dedicado a questões científicas e a prática profissional, voltadas para os problemas da efetiva comunicação do conhecimento e de registros de conhecimentos entre seres-humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. (SARACEVIC, 1996, p.47)

A CI tem como principal característica a sua interdisciplinaridade, e isso se deve as relações que ela possui com outras áreas do conhecimento (psicologia, museologia, sociologia, biblioteconomia, matemática, lógica, informática, economia, filosofia, etc.). Conforme Saracevic (1996, p. 41), a CI:

> [...] é interdisciplinar por natureza e a interdisciplinaridade está longe de acabar, em segundo lugar, a ciência da informação está inexoravelmente conectada à tecnologia da informação, terceiro, a ciência da informação é, juntamente com outros campos, um participante ativo na evolução da sociedade da informação.

É evidente que a CI tem como objeto de estudo a informação, que por sua vez possui competência para gerar conhecimento nos indivíduos, mas para que isso ocorra em um indivíduo, é necessário que ela circule, que seja transmitida, até chegar ao indivíduo, de forma compreensível (BARRETO, 2007). Segundo Le Coadic (1996), informação é o sangue da CI e sua importância apenas é atribuída quando circula livremente, para então atingir o seu objetivo de ser utilizada e gerar resultados através dela.

Atualmente a informação é disponibilizada de diversas formas, o que não torna isso necessariamente um problema, mas o que ocorre é que essa informação é oferecida de forma desorganizada. Preocupados com isso, estudiosos da CI desenvolveram a representação temática da informação, uma ramificação do campo mais amplo denominado Representação da Informação, voltada especificamente para tentar solucionar a questão da desorganização da informação (GUIMARÃES; SALES; GRÁCIO, 2012).

É importante citar que na literatura os termos Representação da Informação e a Organização da Informação (OI) são utilizados por diversos autores para designar a mesma subárea científica, pois possuem propriedades e elementos semelhantes. Portanto, a OI, assim como a Representação da Informação, é tratada como uma subárea da CI e ambas são temas bastantes discutidos atualmente devido as mudanças de paradigmas acerca da propriedade da informação.

A Representação da Informação foi desenvolvida com o objetivo de tratar e organizar a informação, para assim possibilitar a sua recuperação. Inserida na Representação da Informação, a representação temática da informação abrange os processos de catalogação, classificação e indexação, onde todos empenham-se em representar os conteúdos informacionais presentes nos documentos[1]. Dentre esses processos, a indexação (objeto desta pesquisa) se destaca, pois é muito importante para atingir o objetivo final que é a recuperação a informação.

A indexação é o processo mais utilizado e que auxilia na identificação, tratamento e disponibilização da informação. A indexação é uma das etapas da representação temática, é o ato de selecionar ou definir termos (palavras ou expressões) que melhor irão descrever o conteúdo dos documentos para que possa ser recuperado (FUJITA *et al*., 2009). Segundo Lancaster (2004), indexar é um processo complicado e que varia de acordo a um conjunto de fatores, como: indexador, usuários, instituição.

O processo de indexação é responsável pela tradução dos documentos, saindo de sua linguagem natural para uma linguagem documentária (CINTRA; *et al*, 2002, p.39). Esse processo cria uma linguagem entre o sistema e o usuário, capaz de auxiliar no processo de recuperação da informação. Segundo Cintra, as Linguagens Documentárias “são linguagens construídas para a indexação, armazenamento e recuperação da informação, e correspondem a sistema de símbolos destinados a “traduzir” os conteúdos dos documentos” (CINTRA; *et al*, 2002, p. 33). A tradução é realizada através da atribuição de termos que condizem com o documento que está sendo indexado.

A indexação consiste da atribuição de termos ou palavras que melhor descrevem o conteúdo do documento. Essa atribuição é realizada através de uma análise de assunto, que geralmente é realizada por um profissional (indexador), com

[1] Entende-se por documento nesse trabalho, qualquer informação que esteja registrada.

base em metodologias e procedimentos. No processo de indexação existem duas formas de análise de conteúdo semântico: a manual e a automática.

Os processos de indexação manual e automático são, atualmente, bastante estudados por diversos pesquisadores da CI. Alguns estudos desenvolvidos por esses pesquisadores buscam formas de aprimorar a indexação manual através da melhoria na abstração, objetivando tornar a representação do conteúdo mais fidedigna. Infelizmente, a indexação manual não consegue tratar uma grande quantidade de informações com rapidez. Esse problema é descrito por Borges, Maculan e Lima (2008) como a morosidade da indexação manual. Para combater essa morosidade, surgiu a indexação automática.

A indexação é importante para o funcionamento da Recuperação da Informação (RI) (KURAMOTO, 1995), já que ela fornece a descrição dos documentos através de termos que funcionam como ponto de acesso aos mesmos para fins de recuperação.

O termo Recuperação da Informação surgiu através do pesquisador Calvin Moores, em 1950, segundo ele a RI "trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação". (MOORES *apud* FERNEDA, 2003, p. 11). De acordo com Gonzales e Lima (2003), os sistemas de recuperação de informação “[...] tratam essencialmente de indexação, busca e classificação de documentos (textuais), com o objetivo de satisfazer necessidades de informação de seus usuários, expressas através de consultas”. Segundo Saracevic (1999), a RI está diretamente relacionada ao progresso científico da CI e é essencial para a disseminação da informação e geração de conhecimento.

Através da RI, o usuário é capaz de encontrar a informação desejada dentro de uma unidade informacional. Para que isso ocorra, é necessária a realização de procedimentos relacionados a representação temática da informação, armazenamento, organização e acesso. Para Meadows, Boyce e Kraft (2000) a representação da informação e a interpretação e identificação das estruturas e conjuntos dos símbolos são as bases que fundamentam a RI.

O processo de RI ocorre através dos Sistemas de Recuperação da Informação (SRI) (KURAMOTO, 1997). Os SRIs adotam as palavras isoladas como descritores para a indexação de documentos (SOUZA, 2006). Entretanto, apenas as palavras

isoladas não são suficientes para descrever o documento, pois não carregam valor discursivo (BAEZA-YATES; RIBEIRO NETO, 1999).

Basicamente, a indexação automática realizada pelos SRIs consiste em extrair palavras presentes no documento, através de uma análise de texto realizada por computadores, com fim de descrever o seu conteúdo. Entretanto, as unidades extraídas se tratam de palavras isoladas, ou seja, continuam sem valor discursivo e semântico. Devido a isso, diversos pesquisadores passaram a se preocupar mais com a sintaxe e com a semântica presente nos documentos com o intuito de solucionar o problema da falta do valor discursivo e semântico nas palavras extraídas pelos SRIs e outros problemas presentes na linguagem natural, tais como a sinonímia (múltiplas palavras com o mesmo sentido) e a polissemia (multiplicidade de sentidos de uma palavra).

Kuramoto (1995), afirma que, isoladamente, as palavras extraídas de um documento apresentam uma redução no seu valor, pois há a perda da realidade extralinguística do autor. Com isso, é evidente a necessidade de que os descritores atribuídos a um documento contextualizem e representem a informação sem descaracterizá-la. Isto posto, Kuramoto (1995) apresenta os sintagmas nominais (SN) como sendo uma alternativa as palavras isoladas, uma vez que os SNs são considerados descritores mais adequados já que agregam valor semântico.

Os sintagmas são expressões com relações de dependências que estabelecem elos de subordinação entre elementos, os quais também são sintagmas (BORGES; MACULAN; LIMA, 2008). Segundo Martins (2014, p. 42), os sintagmas são um conjunto de elementos (palavras) que possuem uma unidade significativa na oração e mantém dependência e ordem com o núcleo do sintagma.

Os SNs são sintagmas que apresentam na estrutura o núcleo composto por substantivos, ou pronomes, ou palavras substantivadas (PERINI, 1998).

Atualmente, no campo da CI, a utilização dos SNs como recursos para a recuperação da informação já vem sendo estudada por diversos pesquisadores, entre eles se destacam: Kuramoto (1995; 2002), Souza (2005; 2006), Maia (2008), Maia e Souza(2010), Corrêa *et al*. (2011), Lopes (2012), Martins (2014), Silva (2014), Souza e Raghavan (2014), Nascimento (2015).

Mesmo ressaltando o potencial existente na utilização dos SNs quanto à representação e recuperação da informação, é importante frisar que em pesquisas, como a de Kuramoto (1995; 2002) e Souza (2005; 2006), é observado que nem todos

os SNs presentes no documento são passíveis de serem os seus descritores. Portanto, a extração de SNs, realizada automaticamente, não apresentará resultados absolutos que irão descrever o conteúdo informacional do documento. Logo, deve-se esperar que critérios como especificidade e exaustividade, que são aplicados nas palavras-chave da indexação manual, também sejam aplicáveis aos SNs. Com base nisso, Corrêa *et al*. (2011) afirmam que a extração de SNs

> [...] não garante por si só a seleção de bons descritores, sendo necessário que a ferramenta de extração de sintagmas nominais possa fazer a análise dos textos e pontuar os sintagmas nominais com a potencialidade de serem bons descritores [...] (CORRÊA *et al*., 2011, p. 11).

Para descrever melhor a situação onde nem todos os SNs são passíveis de descrever corretamente o documento, o Quadro 1, apresenta SNs (anteriores ao teste de revocação) retirados do título e resumo do artigo “Transferência da Informação: análise para valoração de unidades de conhecimento”, escrito por Plácida L. V. Amorim da Costa Santos e Ricardo César Gonçalves Sant'Ana, artigo esse presente no corpus de Souza (2005).

Quadro 1 – Sintagmas nominais

Sintagmas Nominais		
Transferência da Informação	o valor de conhecimento	os estudos sobre a gestão do conhecimento
análise para valoração de unidades de conhecimento	o conhecimento	a gestão do conhecimento
valoração de unidades de conhecimento	as mais discutidas	Conhecimento disponível
unidades de conhecimento	menos compreendidas questões	esta dificuldade
Conhecimento	o conjunto de conhecimento de uma organização	uma organização
o valor do conhecimento	o conhecimento de uma organização	o mercado
o conhecimento	Parâmetros	algum processo

(**Fonte:** desenvolvido pelo autor.)

Ao observar o Quadro 1, é possível identificar alguns SNs que podem se tornar os descritores do documento, esses foram destacados em vermelho por serem relevantes e terem valor descritivo para os documentos.

Todavia, entre os SNs apresentados, alguns não possuem valor descritivo para o artigo, como os SNs destacados em verde e em preto. Os SNs em verde tratassem de SNs sem valor como descritor ou irrelevantes para a descrição dos documentos, em paralelo, os SNs em preto são aqueles que não possuem nenhum significado, ou seja, SNs vazios. É nesse cenário que o presente trabalho está inserido, pois pretende apresentar uma proposta de normalização dos SNs em termos canônicos, através de critérios e diretrizes que proporcionem o controle de vocabulário e menor dispersão terminológica, afim de melhor descrever o conteúdo informacional presente nos documentos indexados automaticamente.

Ciente da importância da indexação para a recuperação da informação, da dificuldade encontrada para indexar manualmente o grande fluxo informacional atual, a par do interesse do pesquisador quanto a temática de indexação automática, e com a aproximação com pesquisas de Kuramoto (1995, 1999), Corrêa *et al*. (2011), Lapa (2014), Maia (2008), Souza (2005), entre outros, motivaram o desenvolvimento da pesquisa.

Com base na utilização de SNs na indexação automática, **o problema** da presente pesquisa consiste no desenvolvimento de uma proposta de normalização para os SNs em termos canonizados, a fim de que, em conjunto com seu valor semântico e discursivo, se qualifiquem como descritores dos documentos em processos de indexação automática.

A par do problema da pesquisa, a **problemática** é norteada para a reflexão e análise das seguintes indagações: Como selecionar os SNs mais relevantes para serem normalizados? Todos os SNs normalizados são possíveis descritores? A indexação automática com SNs normalizados descreve corretamente os documentos? Os critérios de normalização podem ser usados para detectar e eliminar SNs vazios e SNs irrelevantes?

Em face do problema e da problemática, a presente pesquisa foi desenvolvida com o intuito de apresentar critérios e diretrizes que conduzirão a normalização dos SNs em termos canonizados e sua aplicação no processo de indexação automática.

1.1 Objetivo Geral

Quanto aos objetivos da pesquisa, tem-se como objetivo geral que orientará o presente trabalho:

- Desenvolver uma proposta de normalização de SNs extraídos automaticamente em termos canônicos, considerando a linguagem documentária, seus princípios e conceitos, como base científica.

1.2 Objetivos Específicos

Com base no objetivo geral, foram desenvolvidos os seguintes objetivos específicos:

1. Investigar os processos ligados a indexação automática por SNs;
2. Selecionar manualmente os SNs mais relevantes;
3. Minimizar SNs extensos;
4. Alterar os SNs para aproximar de termos descritores;
5. Comparar os SNs normalizados com termos do TBCI
6. Avaliar o método proposto em um experimento de normalização dos SNs;

A importância e valor dessa pesquisa deve-se ao fato que será possível determinar se através da normalização de SNs para a indexação automática é possível diminuir a dispersão terminológica observada em SNs não normalizados, consequentemente proporcionando um controle do vocabulário.

É evidente, no cenário atual, a importância que os SNs têm adquirido no processo de indexação, pois, como apresentado em argumentos anteriores, o seu valor como descritor para um documento é superior ao de palavras isoladas. Diante dessa conjuntura, a pesquisa aqui desenvolvida se mostra relevante graças ao desenvolvimento de uma proposta onde os SNs poderão ser transformados em termos canônicos através de critérios e diretrizes que poderão ser aplicadas em sistemas de indexação automática por SNs.

O desenvolvimento dessa pesquisa é justificado devido a necessidade de normalizar os SNs antes de indexar os documentos. Através dessa normalização será possível ter controle do vocabulário desenvolvido através da indexação por SNs e assim diminuir a dispersão terminológica que existe nos SNS não normalizados.

A hipótese levantada nessa pesquisa é que através da normalização de sintagmas nominais na indexação automática é possível diminuir a dispersão terminológica e aproximar os SNs de descritores documentais.

A dissertação aborda temas e pesquisas relacionadas a indexação automática com base em SNs, como: o princípio da organização da informação e representação da informação; a indexação manual e os processos realizados; a indexação automática e os SNs; e o processo de indexação automática com SNs (extração, seleção e normalização).

Portanto, a dissertação está estruturada em cinco capítulos, sendo eles:

1. Introdução;
2. Quadro Teórico Conceitual;
3. Metodologia da Pesquisa;
4. Análise dos Resultados.
5. Conclusão

Nesta introdução foram abordados os principais conceitos que serão detalhados nos próximos capítulos, mostrando um pouco o papel da organização da informação na CI, o crescimento informacional, o desenvolvimento da indexação automática, a utilização de SNs no processo de indexação automática e a importância da indexação para os SRI. Ao fim da introdução são apresentados o problema, a problemática, a importância e os objetivos do trabalho.

No segundo capítulo, é abordado a estrutura lógica que orientou a na construção dos componentes teóricos que são discutidos durante a pesquisa (organização da informação, indexação manual, indexação automática, SNs e indexação automática por SNs).

No terceiro capítulo, é apresentada detalhadamente a metodologia utilizada para a extração e seleção dos SNs com valor descritivo presentes no *corpus* da pesquisa e a proposta de normalização dos SNs que foi definida em duas etapas e aplicada nos SNs selecionados.

No quarto capítulo, são discutidos e analisados os resultados obtidos durante o processo de extração e normalização de SNs, apresentando os principais problemas presentes durante ambos os processos e discutindo quais critérios foram relevantes para a proposta de normalização dos SNs.

O quinto capítulo é o último abordado na dissertação e apresenta a conclusão da pesquisa com algumas discussões sobre os processos utilizados durante a normalização dos SNs, contribuições do trabalho, limitações e trabalhos futuros.

2 QUADRO TEÓRICO CONCEITUAL

Para o desenvolvimento do referencial teórico, foi feito um levantamento bibliográfico a respeito dos seguintes campos temáticos: Indexação Manual, Indexação Automática e os SNs no âmbito da indexação automática.

Entretanto abordar a Organização da Informação é pertinente para o trabalho, pois inserida nesse campo está a Representação Temática da Informação (RTI), que tem como uma de suas premissas a investigação da ligação estabelecida pela informação documentária entre documentos e usuários.

Vinculados a Organização da Informação, temos os processos de indexação que, por sua vez, se apresentam de formas diferentes devido ao modo que são praticados. Logo, é realizada uma revisão da Indexação Manual e da Indexação Automática, apresentando suas características, conceitos, peculiaridades e estudos quanto a sua aplicação.

Posteriormente, é feito um levantamento sobre os SNs e a sua relação com a indexação automática, buscando compreender seus conceitos, o valor, as contribuições e as relações que os SNs podem oferecer no processo de indexação automática, afim de que, através desse levantamento, possa ser construído o principal objeto dessa pesquisa, que é a proposta de normalização.

2.1 Organização da Informação

Segundo Feitosa (2006), no que tange a temática de tratamento da informação, é importante compreendermos os conceitos básicos da área da CI (CI). Portanto, para discutir sobre o processo de Organização da Informação (OI), é necessário que retomemos ao conceito básico do que é informação. Logo, a informação no âmbito da CI, segundo Le Coadic (2004, p. 4) é:

> um conhecimento inscrito (registrado) em forma escrita (impressa ou digital), oral ou audiovisual, em um suporte. A informação comporta um elemento de sentido. É um significado transmitido a um ser consciente por meio de uma mensagem inscrita em um suporte espacial-temporal: impresso, sinal elétrico, onda sonora, etc. inscrição feita graças a um sistema de signos (a linguagem), signo este que é um elemento da linguagem que associa um significante a um significado: signo alfabético, palavra, sinal de pontuação. (LE COADIC, 2004, p. 4)

Percebe-se que a informação está relacionada ao conhecimento, lembrando que para a informação se tornar conhecimento é necessário que ela seja utilizada, ou seja, recuperada. Todavia, para que ela possa ser recuperada e utilizada, é imprescindível que ela esteja organizada; nesse momento, surge a OI.

Segundo Dias (2006), a OI é um processo que trata questões relacionadas a descrição física e do conteúdo do documento, e tem como produto a representação documental, como: os registros bibliográficos e os índices.

É importante ressaltar que a OI não é uma preocupação recente, pelo contrário, segundo Robredo (2004), existem evidências de que povos da Mesopotâmia (há mais de 4000 anos) tinham controle e faziam registros organizados em tabuletas de argila. Entretanto, foi no século XIX que passou a se investigar mais sobre essas técnicas, e isso ocorreu devido ao crescimento da produção bibliográfica, da pesquisa científica e do surgimento das profissões de Arquivologia e Biblioteconomia, que estudam os registros da informação. De acordo Shera e Egan (1961), no século XX surgiu a oportunidade de desenvolver e experimentar novos métodos de organização "[...] especialmente durante e depois da Segunda Guerra Mundial, a fim de satisfazer às modernas exigências da pesquisa bibliográfica intensiva e aprofundada". Medeiros (2010) ratifica essa discussão ao afirmar que a Documentação e a Biblioteconomia deram origem a OI.

Ao analisar estudos referentes a OI, percebe-se que ela também é chamada de Organização do Conhecimento (OC) por alguns autores (BRASCHER; CAFÉ, 2008). Apesar dessa divergência, a OI e a OC possuem assuntos que são interligados (RIBEIRO; NAVES; KURAMOTO, 2006; BRASCHER; CAFÉ, 2008).

> Organização do conhecimento está relacionada com um processo de análise conceitual de um domínio do conhecimento, e, a partir daí sua estruturação, gerando uma representação do conhecimento de tal domínio. Dessa forma, obtém-se um instrumento - um esquema de representação do conhecimento - que será então usado para a organização da informação desse domínio de conhecimento produzida. (BRANDT; MEDEIROS, 2010, p. 112)

Burke (2003), ao tratar sobre o que seria conhecimento, afirma que é necessário discernirmos sobre os conceitos de informação e conhecimento. Segundo o autor, a informação se refere a algo "cru", e o conhecimento se refere a algo "cozido",

ou seja, a informação seria a matéria bruta para o desenvolvimento e construção de um conhecimento. Por outro lado, alguns autores afirmam que a informação advém do conhecimento (FOGL, 1979 *apud* BRÄSCHER; CAFÉ, 2008).

Bräscher e Café (2008, p. 4) declaram que ambos os conceitos, informação e conhecimento, apesar de sua relação, possuem características que os distinguem e com isso é possível delimitar a utilização dos termos de acordo com a necessidade. Ademais, afirmam existir dois processos de organização dispares. Um deles é voltado aos conceitos, tendo como unidade de pensamento o processo de OC. O outro é vinculado ao objeto informação, tendo como unidade de pensamento a OI. Dentro desta discussão está a Representação Temática da Informação (que é um produto da OI) e a Representação do Conhecimento (RC) (resultado da OC).

A palavra representação, etimologicamente, é originária do latim "*repraesentatio*", de "*repraesentare*". Seu conceito remete a domínios do conhecimento. No âmbito filosófico, representar é a presença, em espírito, do objetivo, e na psicologia, se trata da imagem mental referente a um objeto ou a uma situação. Contudo, o termo representação está presente desde a pré-história. Pinto *et al* (2008), afirmam:

> O significado que a palavra representação encerra não é de origem tão recente, conforme parecem imaginar alguns. Muito pelo contrário, ela sempre esteve presente no espírito humano, pelo menos, desde a Pré-história quando os homens primitivos, em suas práticas cotidianas, buscavam possibilidades de comunicação através da criação de imagens ou ideogramas; assim como da escrita cuneiforme dos sumérios e dos hieróglifos produzidos no Antigo Egito.

No campo científico da CI, Pinto e Meunir (2006), afirmam que representar é compreendido como:

> […] ação de construir etiquetas (labels ou tags) mentais utilizadas para indicar ou apontar as coisas do mundo, através dos signos verbais ou não verbais. Ou seja, estas etiquetas consistem no objeto representante que vai estar no lugar do objeto representado, para dar sentidos a ele, a fim de facilitar a compreensão do mundo e a comunicação entre os seres. (PINTO; MEUNIER, 2006).

Na OI existem processos responsáveis por preparar a informação para utilização em Bibliotecas e Sistemas de Recuperação de Informação (SRIs). Esse conjunto de processos compõem o "Tratamento da Informação", que, por sua vez, se divide em dois tipos: a representação descritiva e a representação temática.

A representação descritiva é responsável por descrever as informações físicas do recurso. A representação temática está relacionada à representação do conteúdo informacional, à temática e ao assunto abordado no documento. Além disso, a representação temática oferece processos como classificação, indexação e catalogação que são utilizados na descrição do documento (DIAS e NAVES, 2007; GUIMARÃES, 2009; REDIGOLO, 2010). Ambas as representações são muito importantes para a OI, pois é através delas que a informação chegará ao usuário final.

Apesar da representação temática e descritiva serem processos distintos, existe uma relação entre elas, pois ambas realizam atividades que possibilitam a organização e a propagação da informação. Ruiz Perez (1992) enuncia que essas atividades são: a análise documental de forma e a análise documental de conteúdo.

Corroborando com o que foi apresentado até o momento, Dias e Naves (2007) afirmam que

> Podemos, portanto, sintetizar o conceito de tratamento da informação da seguinte forma: expressão que engloba todas as disciplinas, técnicas, métodos e processos relativos a: a) descrição física e temática dos documentos numa biblioteca ou sistema de recuperação de informação; b) desenvolvimento de instrumentos (códigos, linguagens, normas, padrões) a serem utilizados nessas descrições; e c) concepção/implantação de estruturas físicas ou bases de dados destinadas ao armazenamento dos documentos e de seus simulacros (fichas, registros eletrônicos, etc.).

O processo de representação se divide em dois níveis, o primário e o secundário. O primário é compreendido como a representação realizada pelo autor no momento em que está registrando suas ideias e pensamentos no documento. Alvarenga (2003) relata a existência de algumas etapas realizadas pelo autor no processo de registro do conhecimento, que são: a interpretação, a identificação, a reflexão, a percepção, e a codificação.

No nível secundário de representação é onde ocorre o processo de análise documental, responsável pela representação do conhecimento presente nos documentos primários, através da extração de conceitos representativos, como pontos de acesso. Alvarenga (2003) afirma que durante o processo de tratamento e organização da informação são utilizadas diferentes ferramentas e técnicas que permitem aos profissionais da informação desenvolverem diferentes tipos de representação, transformando a informação primária em registros específicos, objetivando sua recuperação.

Através da análise documental é possível obter uma representação documentária, mas para que isso ocorra é necessária a aplicação de diversos procedimentos e critérios, como a padronização e a univocidade. Segundo Pereira e Bufrem (2005), dois critérios são essenciais qualificadores das principais formas de representação: o resumo e o índice.

Verifica-se que quando discutida a Representação Temática da Informação (RTI), existe uma incerteza quanto a sua terminologia. É comum a presença de sinônimos que, por sua vez, não são necessários, pois tem a mesma finalidade e apenas ocasionam um desconforto para a área científica. Como exemplo, foram observadas as diversas terminologias existentes para a “Análise de Assunto”, que, por sua vez, se apresenta, em diversas outras obras, como “Análise de Informações”, “Análise Documentária”, e “Análise Conceitual”.

Segundo Guimarães (2009), o tratamento temático da informação (TTI) tem seu desenvolvimento norteado por duas vertentes teóricas: a baseada em concepções filosóficas (advindas de Platão, Aristóteles, Bacon e outros) relativas a divisão do conhecimento, e a vertente pragmática, baseada na necessidade prática de organização documental.

Inseridas no universo da OI, as atividades realizadas para a representação da informação: catalogação, indexação e classificação, são norteadas por três correntes teóricas: a Norte-Americana, a Inglesa, e a Francesa.

A abordagem Norte-Americana é responsável pela Catalogação de Assunto (*Subject Cataloging*), enquanto que a Inglesa é responsável pela Indexação (*Indexing*), e Francesa pela Análise Documentária (*Analyse Documentaire*) (GUIMARÃES, 2009, p. 106).

A corrente teórica Norte-Americana foi, por muitas vezes, guiada pelos princípios presentes na catalogação alfabética de Cutter e pelos cabeçalhos de assunto da *Library of Congress* (LC). Além disso, possui uma abordagem focada no pragmatismo e está voltada ao desenvolvimento de produtos através da RTI.

Segundo Sanchez Luna (2004, p. 83), a catalogação é uma “operação pela qual se identifica o documento em função de suas características formais e de seu conteúdo, tais como o autor, o título, o local de publicação, o editor, o ano de publicação, assim como o tema da obra”. Logo, o processo de catalogação aborda a descrição da forma e do conteúdo do documento, utilizando a catalogação descritiva, a catalogação de assunto e a classificação. Segundo a autora, a catalogação é

conduzida por um conjunto de normas estritas, com o objetivo de preparar a informação para registros bibliográficos.

Coates (1988, p. 10) afirma que a catalogação de assunto tem a função de dirigir o usuário, através de palavras-chave, ao documento. Além disso, ele destaca a possibilidade de o usuário identificar documentos através de determinado assunto ou assuntos correlacionados.

Alguns autores afirmam que através da catalogação de assunto é desenvolvida a "análise de assunto", e ela é responsável por gerar a representação do conteúdo informacional, através de notações classificatórias ou de cabeçalhos de assunto (RAJU & RAJU, 2006, p. 15). Porém, Sauperl (2002) ressalta que a análise de assunto se trata de um processo anterior a catalogação de assunto, pois desenvolve os fundamentos que possibilitam a tradução da informação e a construção de uma linguagem de classificação ou de um vocabulário controlado.

A corrente teórica Inglesa (norteadora do desenvolvimento dessa pesquisa) foi a responsável pela indexação, tendo em sua concepção a função de apoiar a pesquisa das bibliotecas especializadas (GUIMARÃES, 2009). Ademais, alguns autores, como Neet (1989), afirmam que a indexação possui similaridades com a catalogação de assuntos. A indexação é uma atividade com grande valor para unidades e sistemas de informação, pois trata a informação presente em documentos, abordando não apenas o documento em si, mas a informação contida nele, apresentando outra dimensão informacional.

A utilização do termo dimensão informacional se deve ao fato de que a indexação trata a informação através de duas dimensões. Segundo Neet (1989, p. 7) a indexação consiste em "facilitar a pesquisa de documentos ou de informações contidas em documentos", logo estão presentes duas dimensões: documento e informação.

A indexação apresenta uma preocupação, com relação aos usuários, quanto ao uso e recuperação da informação. Apoiado nessa premissa, a indexação passou a moldar os sistemas de informação almejando uma harmonia com os hábitos de busca dos seus usuários, ao invés de forçá-los a se adaptarem aos moldes de busca do seu sistema. Fidel (2000, p. 79-80) mostra que os sistemas de informação devem estar de acordo com as necessidades dos usuários e não segundo regras universais, pois quanto mais um sistema busca adequar-se às necessidades e ao comportamento do usuário, mais amigável será o uso e a recuperação da informação.

É sabido que nas correntes teóricas inglesa e norte-americana os estudos buscavam desenvolver metodologias para o desenvolvimento de produtos de tratamento de informação. Por outro lado, a corrente francesa surgiu durante o início da década de 70 e tinha como foco compreender o processo de tratamento da informação afim de desenvolver referenciais teóricos-metodológicos, se diferenciando das correntes Norte-Americana (catálogos) e Inglesa (índices), que tinham como foco desenvolver produtos.

Jean-Claude Gardin (1966) e Coyaud (1966), através de seus trabalhos, desenvolveram a Análise Documental, um estudo linguístico que abrange a indexação. Segundo Chaumier (1982, p. 27) *apud* Guimarães (2009), "a análise documental abrange dois tipos de tratamentos diferentes: a condensação, que se vale de uma redução do texto para fins de difusão da informação, e a indexação, que se vale da extração de conceitos para servir de apoio à recuperação".

2.2 Indexação Manual

A indexação é uma atividade que realiza uma seleção de termos (palavras-chave ou expressões) que farão a descrição do conteúdo informacional de um documento, ou seja, é a representação, abreviada, do conteúdo, considerando sempre as necessidades do usuário. Souza e Fujita (2014, p. 22), afirmam que:

> O processo de indexação, além de ter foco no que é abordado no documento, também deve ser direcionado para a necessidade de informação do usuário, materializada por ele na forma de pergunta. É um processo com duas direções: um lado os dos documentos e de outro, as necessidades de informação dos usuários

No que tange o processo de indexação, verifica-se que a seleção dos termos que farão a descrição do conteúdo também é chamada de tradução, porém, no decorrer desse processo de tradução, existem possíveis problemas quando relacionados a sinonímia e homonímia.

Segundo o Sistema Mundial de Informação Científica (UNISIST) (1981), a representação dos conteúdos presentes em um documento, através de termos (palavras ou descritores), é denominada indexação. O processo de indexação é uma das atividades mais complexas dos bibliotecários.

Para a realização da indexação é necessário um profissional com a formação especializada para o tratamento de documentos. Para Naves (2004):

> O profissional da informação que desenvolve a atividade de indexar assuntos de documentos é chamado de indexador, catalogador de assuntos ou classificador. A maioria desses profissionais é graduada em Biblioteconomia, e deve conhecer os fundamentos teóricos e técnicos do tratamento temático da informação. (NAVES, 2004)

Todavia, diversos autores alegam que não há uma definição acerca de qual a formação do profissional indexador, permitindo, então, que outros profissionais como, engenheiros, arquitetos, médicos, e outros, possam realizar essa atividade (BARBOSA, 1998).

Segundo Dias e Naves (2007), no âmbito do tratamento da informação, a indexação é compreendida em dois sentidos: amplo e restrito. O sentido amplo está relacionado a criação de índices (autor, título, assunto, livros, periódicos, etc.), catálogos, e banco de dados para bibliotecas ou unidades de informação. O sentido restrito é focado, apenas, na catalogação de assuntos referentes as informações contidas em documentos. Farrow (1995 *apud* DIAS; NAVES, 2007, p. 27), distingue esses sentidos, ao afirmar que:

> indexação back-of-book permite ao leitor localizar informação sobre um tópico dentro do livro; a tarefa do indexador é ler o texto, distinguir entre informação relevante e periférica e empregar juntos o processamento top-down (conceitual) e bottom-up, presentes na leitura fluente. Por sua vez, a indexação acadêmica fornece um termo útil estabelecido pela indexação praticada em bases de dados de resumos e em catálogos de bibliotecas, usando predominantemente a abordagem top-down. A segunda é considerada menos exaustiva que a primeira. (FARROW, 1995 *apud* DIAS; NAVES, 2007, p. 27)

O processo de indexação de assuntos pode apresentar diversas dificuldades. Isso varia de acordo com o assunto abordado pelo documento, se o assunto for bastante complexo ou implícito, é esperado que o nível de dificuldade seja maior (BERNIER, 1965). Profissionais em recuperação da informação afirmam que indicar termos com valor descritivo do conteúdo de um documento é a atividade mais importante e também a mais difícil, já que envolve diversas variáveis, e seu resultado interfere diretamente na recuperação da informação (SALTON. McGILL, 1983).

Considerando o fato da indexação ser uma atividade intelectual, é comum percebermos divergências entre os termos que representam um mesmo documento, haja vista que podem ter sido indexados por profissionais distintos e, possivelmente, em contextos diferentes. Além disso, para Lancaster (2004), determinado documento pode apresentar um conjunto de termos diferentes de indexação, variando de acordo

com o grupo de usuários ao qual o documento tratado será destinado, ou seja, é possível que o mesmo documento apresente formas diferentes de indexação, mas que ambas estejam corretas.

Consequentemente, verifica-se na literatura que o processo de indexação, com relação ao desenvolvimento de diretrizes e critérios quanto a utilização de linguagens documentais, é, às vezes, realizado de maneira superficial ou ingênua, sem analisar as possíveis variações linguísticas e lógicas que acompanham essa atividade. Isso é destacado com a afirmação apresentada por Cleveland & Cleveland (1990, p. 136, *apud*. GUIMARÃES, 2009, p. 109):

> Como em qualquer processo de indexação de assunto, o indexador começa com o reconhecimento das próprias palavras do texto, “escaneando” cada sentença e grifando as palavras-chave utilizadas pelo autor. O indexador, então, avalia as referidas palavras face à estrutura geral do parágrafo de modo a determinar os assuntos que estão sendo discutidos. Certamente, nem toda palavra grifada em um parágrafo constitui assunto significativo. Muitas palavras são de menor importância e não seriam utilizadas em um índice. Que tópico ou tópicos importantes são discutidos no parágrafo? Que palavras são simplesmente modificadores e não efetivamente indicadores de assunto? Por exemplo, em os mosquitos atacaram com a ferocidade de um tigre, apenas mosquitos são um indicador de assunto.

Diversos especialistas afirmam que além do processo de indexação ser uma atividade cognitiva, ele é acompanhado de uma certa subjetividade, a qual contribui para a existência de diferentes indexações. Strehl (1998) afirma que essa subjetividade se deve aos múltiplos julgamentos, aos níveis de concordância e as disparidades geradas durante o processo de indexação. Diversos fatores colaboram para a existência dessas diferenças na indexação (DIAS; NAVES, 2007), tais como:

- Diferentes indexadores;
- Momentos de indexação;
- Conteúdo do documento;
- Conceitos importantes para a representação do conteúdo;
- Que parte do conteúdo realmente responde as necessidades dos usuários;
- Os descritores definidos para representar esses conceitos (STREHL, 1998)

Para Silva (2006) e Fujita (2007), o indexador é influenciado por outros elementos, além da política de indexação, e também o contexto onde está inserido.

Tal contexto engloba três aspectos: Aspecto Físico, Aspecto Psicológico e Aspecto Sociocognitivo.

- Aspecto físico: está relacionado com a biblioteca em si, sua estrutura física, e os documentos que estão sob os cuidados do indexador;
- Aspecto psicológico: engloba os problemas que, por ventura, o indexador possa estar passando;
- Aspecto sociocognitivo: inclui a política de indexação, os objetivos da indexação, regras e procedimentos, a linguagem documentária, a mediação da linguagem dos usuários e os interesses de busca do usuário.

É importante salientar que durante a indexação podem estar presentes problemas relacionados a linguagem natural como: polissemia, sinonímia e combinações diferentes palavras que geram significados diferentes (NAVARRO, 1988).

Segundo Vieira (1988) a indexação é um dos processos básicos de recuperação da informação, podendo então ser feita pelo homem (indexação manual) ou por computador (indexação automática).

Portanto, é evidente uma relação direta entre a indexação e a recuperação da informação, logo é importante a existência de uma indexação normatizada através de uma política de indexação (LANCASTER,1968).

A UNISIST elaborou um documento que definia os princípios da indexação referentes a indexação manual. A criação desse documento é considerada a primeira tentativa internacional de normalizar o processo de indexação.

Durante o processo de indexação manual, os termos que descrevem o documento são extraídos através de análise intelectual, que compreende três fases:

1. Interpretação do conteúdo através da leitura dos componentes do documento, tais como o texto completo, o título, o resumo e outros. A UNISIST recomenda não se focar apenas no título e no resumo, já que nem todos possuem informações suficientes para a extração de termos que irão descrever o documento.
2. Identificação dos conceitos que irão descrever o documento.
3. Seleção dos conceitos, sempre considerando fatores como: exaustividade, especificidade e consistência.

Além da UNISIST, outros autores afirmam que o processo de indexação é composto por etapas, variando de duas até oito (Quadro 2). Lancaster (2004) e Chaumier (1986) afirmam ser apenas duas etapas; a Associação Brasileira de Normas Técnicas (ABNT) cita três; Robredo (2005), Chu & O'Brien (1993) e Van Slype (1991) citam quatro; por fim, Guinchat e Menou (1994, p. 177) apresentam oito etapas (RUBI, 2008).

Quadro 2 – Etapas da indexação

Autores	Etapas
Lancaster (2004) e Chaumier (1986)	1. Análise Conceitual; 2. Tradução
ABNT (NBR 12676/1992)	1. Exame do documento e estabelecimento do assunto de seu documento; 2. Identificação dos conceitos presentes no assunto; 3. Tradução desses conceitos nos termos de uma linguagem de indexação.
Robredo (2005), Chu & O'Brien (1993) e Van Slype, (1991)	1. Análise de Assunto do texto; 2. Expressão do conteúdo do assunto nas palavras dos indexadores (Linguagem Natural); 3. Tradução para um vocabulário de indexação; 4. Expressão do assunto em termos do índice.
Guinchat e Menou (1994, p. 177)	1. Lembrar os objetivos da operação, se for o caso; 2. Tomar conhecimento do documento; 3. Determinar o assunto principal do documento; 4. Identificar os elementos do conteúdo que devem ser descritos e extrair os termos correspondentes; 5. Verificar a pertinência dos termos escolhidos; 6. Traduzir os termos da linguagem natural nos termos correspondentes da linguagem documental, se for o caso; 7. Verificar a pertinência da descrição; 8. Formalizar a descrição se o sistema prevê regras particulares de apresentação ou de escrita.

(**Fonte:** Desenvolvido pelo autor).

Silva e Fujita (2004, p. 136-137) afirmam que "o conceito de indexação surgiu a partir da elaboração de índices e atualmente está mais vinculada ao conceito de análise de assunto". Esse vínculo é notado no Quadro 2, onde, nas etapas descritas por cada um dos autores, está presente a análise de assunto. Ademais, as autoras consideram que a indexação se trata de uma operação de tratamento temático, visto que são praticadas as atividades de análise, síntese e representação do conteúdo do documento.

Com base nas etapas apresentadas e na afirmativa das autoras Silva e Fujita (2004), ao longo desse trabalho serão consideradas como principais fatores e norteadores da indexação as etapas de análise de assunto (ou análise conceitual) e a tradução, visto que, para o desenvolvimento da ferramenta que será proposta nesse trabalho, essas etapas são essenciais.

2.2.1 Análise de Assunto

A análise de assunto identifica qual assunto é abordado no documento. Para isso, é necessário que seja feita uma leitura almejando compreender, a fundo, quais assuntos abordados são primordiais. Dias e Naves (2007, p. 20) afirmam que para realizar a análise de assunto “é preciso que seja feita uma leitura que possibilite a extração de conceitos que sintetizem o conteúdo desses textos”. Todavia, as autoras enfatizam que “o tempo que pode ser dedicado ao estudo e à reflexão sobre o próprio ato de análise de assunto é provavelmente pequeno, porque à espera do indexador estará certamente, quase sempre, um grande volume de documentos para serem lidos e analisados”. Portanto, deve ser feita uma leitura técnica.

> Um documento, inserido num SRI, antes de ser lido pelo leitor, usuário final do sistema, é lido por um leitor técnico, o indexador, aquele que faz a leitura para fins documentários. Esse tipo de leitura, conhecido como leitura documentária ou leitura técnica, tem certas características, não sendo realizada para lazer ou aprendizagem, nem é prazerosa, muito pelo contrário. O alto grau de incerteza, ansiedade e responsabilidade contido na atividade já mostra que a mesma traz pouca satisfação. É um tipo de leitura bem racional e rápido, em que o leitor técnico não tem chances de aproveitar a leitura, já que seu propósito é o de extrair o conteúdo informativo do texto, tendo em vista a sua posterior recuperação por um leitor interessado. (NAVES, 2004, p. 7)

Segundo a UNISIST (1981), para compreender o conteúdo de um documento é necessária uma leitura completa. Entretanto, a prática dessa leitura, atualmente, não é viável, devido ao grande volume informacional. Com base nessa premissa, é orientada a leitura dos seguintes componentes dos documentos, pois são considerados os mais importantes para o indexador:

- Título;
- Introdução e as primeiras frases de capítulos e parágrafos;
- Ilustrações;
- Tabelas;
- Diagrama e suas explicações;
- Conclusão;
- Palavras sublinhadas ou impressas diferente (destacadas).

Cesarino e Pinto (1980) afirma que a análise de assunto é uma etapa com grande importância para a recuperação da informação, uma vez que os SRI se utilizam dessa etapa para inserir dados que irão auxiliar o usuário na busca pela informação.

De acordo com Fujita (2003, p. 64), a análise de assunto pode ser dividida em três procedimentos: compreensão do conteúdo, identificação dos conceitos, e seleção de conceitos para recuperação. Semelhante aos procedimentos propostos por Fujita (2003), Dias e Naves (2007) apresentam os seguintes procedimentos para a análise de assunto: leitura técnica do documento, extração dos conceitos, e determinação da atinência[2].

Segundo Naves (2004, p. 6), a compreensão do conteúdo, por parte do leitor, se dá através do processamento que ocorre em sua mente durante a leitura, e esse processamento acontece de maneira interativa, determinando ou não a compreensão de um texto. Além disso, a autora diz que, com relação ao indexador, a compreensão do conteúdo passa por dois tipos de processamentos mentais da informação:

> [...] o *top-down* e o *bottom-up*, que parecem ocorrer simultaneamente na mente humana ao fazer a leitura de um texto. São inversos e complementares, e chamados por alguns autores de modelos de leitura: é o tipo ascendente, guiado por dados, indutivo, *bottom-up*, no qual a leitura é linear, das partes para o todo textual, e o tipo descendente, dedutivo, *top-down*, no qual se move na forma inversa, obtendo vantagem da base de conhecimento do leitor. Trata-se de uma dupla ação: percepção e compreensão. (NAVES, 2004, p. 6)

Após a realização do processo de compreensão do conteúdo, é necessário realizar a Extração de Conceitos. Conceitos são unidades do conhecimento representados por um termo ou palavra, são obtidos através de enunciados retirados de um documento, e permitem representar ou categorizar o documento. Termos são unidades mínimas da terminologia e correspondem a um conceito em uma linguagem de especialidade (ISSO 704; ISO 1087-1). Entre os conceitos é possível identificar algumas relações, que podem ser de cunho: Hierárquico, Equivalente ou Associativo.

No processo de extração existem duas variáveis que afetam na escolha dos conceitos: a exaustividade e a especificidade. A exaustividade está relacionada a quantidade de termos que foram escolhidos como descritores do assunto do documento, ou seja, quanto mais exaustiva a indexação, mais termos serão atribuídos ao documento. A especificidade se trata do nível de revocação dos conceitos escolhidos para representar o assunto. Ambas as variáveis influenciam no processo de recuperação da informação.

[2] Termo traduzido do "*aboutness*", para indicar qual o conteúdo abordado em um documento.

Tanto a especificidade como a exaustividade estão diretamente ligadas a revocação e a precisão da indexação. Quando uma indexação é feita mais especifica os resultados de recuperação será de uma revocação menor e de uma precisão maior. Por sua vez a exaustividade é o contrário, quando mais exaustiva uma indexação menor será a precisão do sistema, mas será maior a revocação.

Ao finalizar o processo de extração, temos, o que muitos consideram ser o processo mais importante para a análise de assunto, a seleção dos conceitos. É importante mencionar que nem todos os conceitos extraídos durante o processo de extração são necessariamente selecionados para representar o assunto do documento. Para que a seleção de conceitos seja feita da forma mais adequada possível é necessário conhecer bem quais são os assuntos abordados pelo documento para então relacioná-los com os conceitos que melhor os representam.

É importante frisar que a prática da análise de assunto é bastante subjetiva, complexa e é influenciada por diversos fatores, como: necessidades do usuário, SRI, entre outras. De acordo com Naves (2004, p. 10), a análise de assunto é a etapa da indexação que sofre mais influência da subjetividade do indexador.

2.2.2 Tradução

Após a análise de assunto é dado início a etapa de Tradução. Nesse momento, o indexador é responsável por “traduzir” os conceitos, escolhidos na análise de assunto, para uma linguagem de indexação. A linguagem de indexação se subdivide em dois tipos: Linguagem Natural (LN) (Folksonomia) e Linguagem Documentária (LD) (tesauros, cabeçalhos de assunto e esquemas de classificação).

Segundo Lopes (2002, p. 48), a linguagem natural é “[...] sinônimo de discurso comum, isto é, a linguagem usada habitualmente na fala e na escrita sendo que, nas bases de dados, os termos do título e resumo representam a LN”.

É comum alguns indexadores utilizarem a linguagem natural[3] (termos sem alterações extraídos do documento) como uma linguagem de indexação, mas não é o recomendado, pois outros indexadores podem usar outros termos que expressam o mesmo conceito (dispersão terminológica), ou o mesmo conceito pode ser expressado por termos diferentes (dispersão sintática). De acordo com Fujita e Rubi (2006, p. 52),

[3] De acordo com Lancaster (2004, p. 250), a linguagem natural é “utilizada habitualmente na escrita e na fala, e que é o contrário de vocabulário controlado”

> [...] a linguagem de indexação afeta o desempenho de um sistema de recuperação de informação tanto na estratégia de busca (estabelece a precisão com que o técnico de busca pode descrever os interesses do usuário) quanto na indexação (estabelece a precisão com que o indexador pode descrever o assunto do documento). Portanto, a partir de estudos do sistema, deve-se optar entre linguagem livre ou linguagem controlada e linguagem pré-coordenada ou pós-coordenada. (FUJITA; RUBI, 2006, p. 52)

Lopes (2002) destaca algumas desvantagens e vantagens na utilização da LN como uma linguagem de indexação (Quadro 3).

Quadro 3 – Vantagens e desvantagens da linguagem natural como linguagem de indexação

Vantagens	Desvantagens
Permite o imediato registro da informação em uma base de dados, sem necessidade de consulta a uma linguagem de controle.	Os usuários da informação, no processo de busca, precisam fazer um esforço intelectual maior para identificar os sinônimos, as grafias alternativas, os homônimos, etc.
Processo de busca é facilitado com a ausência de treinamentos específicos no uso de uma linguagem de controle.	Haverá alta incidência de respostas negativas ou de relações incorretas entre os termos usados na busca (por ausência de padronização).
Termos de entrada de dados são extraídos diretamente dos documentos que vão constituir a base de dados.	Custos de acesso tendem a aumentar com a entrada de termos de busca aleatórios.
Temas específicos citados nos documentos podem ser encontrados.	Uma estratégia de busca que arrole todos os principais conceitos e seus sinônimos deve ser elaborada para cada base de dados (ex: nomes comerciais de substâncias químicas não ocorrem no *Chemical Abstracts*).
Elimina os conflitos de comunicação entre os indexadores e os usuários, pois ambos terão acesso aos mesmos termos.	Perda de confiança do usuário em uma possível resposta negativa.

(**Fonte:** LOPES, 2002)

Segundo Vale (1987), a escolha da linguagem de indexação é de grande importância, pois ela define o nível de eficácia do SRI, logo essa escolha deve levar em consideração: o usuário, o objetivo do sistema e a revocação ou especificidade dos assuntos abordados. Conhecer o perfil do usuário que irá buscar a informação presente no sistema é muito importante. É preciso identificar seu nível de

conhecimento, para então escolher a linguagem de indexação mais adequada, fazendo com que suas necessidades informacionais sejam atendidas.

As linguagens documentárias foram desenvolvidas com base na linguagem natural, logo possui semelhanças e diferenças. O uso da LN não necessita de conhecimento prévio das regras que a regem, diferente da LD, que trata de um sistema de relações desenvolvido para um determinado universo temático.

Muitas vezes a escolha da LD deve-se ao desejo de melhorar a eficácia do SRI, pois através dela é possível reduzir a redundância, a ambiguidade, a polissemia, e as variações dos elementos, conquistando assim uma univocidade interpretativa.

A diferença entre a LN e a LD está na finalidade e na função que desempenham. A LN é utilizada para diversos fins e tem função descritiva ou factual, conativa, prescritiva, expressiva, evocativa ou estética, fática ou social, e de meta linguagem. A LD possui algumas funções específicas, uma delas é possibilitar a comunicação do leitor com o documento. No eixo sintagmático, tem como objetivo a indexação e a busca; no eixo paradigmático, tem a função de guiar o indexador na escolha dos descritores do documento e orientar os usuários na escolha dos descritores que irão compor a busca. Para se desenvolver uma LD, é preciso estabelecer os níveis e as funções que serão desempenhadas, de forma clara e estruturada (HUTCHINS, 1975, p. 8-11).

É possível estabelecer dois níveis de diferenciação entre a LN e a LD: nível formal e nível semântico. Na LN, no nível formal, a forma escrita é secundária em relação à forma vocal e, no nível semântico, temos a ocorrência de sinonímia e homonímia. Na LD, no nível normal, a forma escrita corresponde a sintagmas de símbolos notacionais (números, letras e pontuação) e, no nível semântico, é almejada uma univocidade entre os descritores e os sintagmas (HUTCHINS, 1975).

Navarro (1998, p.54-55) apresenta um quadro comparativo entre a LN e a LD.

Quadro 4 – Linguagem natural e Linguagem documentária

Linguagem Natural	Linguagem Documentária
São gerais	São especializadas
São estabelecidas e adaptadas através de longos períodos de tempo e por milhares de pessoas	São estabelecidas em poucos anos por um número reduzido de pessoas
São naturalmente aceitas e adquiridas pelos usuários	Devem ser aceitas pelos usuários
São naturais	São artificiais
Têm sua própria estrutura	Sua estrutura baseia-se na estrutura da linguagem natural sobre a qual elas são formadas
São menos eficientes que as Linguagens Documentárias nas operações de recuperação de informação	São mais eficientes que a linguagem natural nas operações de recuperação de informação
São sensíveis a mudanças culturais	São sensíveis a mudanças culturais
Caracterizam-se pela dupla articulação	Não abrangem o conceito de dupla articulação
Têm sua própria teoria	Baseiam-se na teoria das ciências, da ciência da informação e da linguística
Compreendem a noção de morfema e lexema	Compreendem a noção de informema
Não têm funções específicas [podem ter várias]	Têm um propósito específico e um nível de funções
Necessitam respeitar uma hierarquia de traços para evitar malformações gramaticais ou atenuá-las	Necessitam de hierarquias semânticas e sintáticas para evitar malformações
As funções conativa, emotiva, fática e poética (entre outras já mencionadas) são próprias da linguagem natural	Não são dotadas das funções conativa (imperativa), emotiva (interjeição), fática (mensagem que serve para estabelecer, prolongar ou interromper a comunicação) e poética
Os monemas autônomos e funcionais assim como as modalidades, são elementos da linguagem natural	Não comportam pronomes nem modalidades como o artigo, o número, o tempo e pessoa nem categorias como o advérbio e adjetivos

(**Fonte:** NAVARRO, 1998, p. 54-55.)

A linguagem documentária pode ser dividida em duas linguagens: pré-coordenada e pós-coordenada.

A linguagem pré-coordenada é aquela que combina os termos no momento da indexação e os assuntos são representados por um conjunto de termos já combinados, mas, infelizmente, não consegue prever todas as combinações.

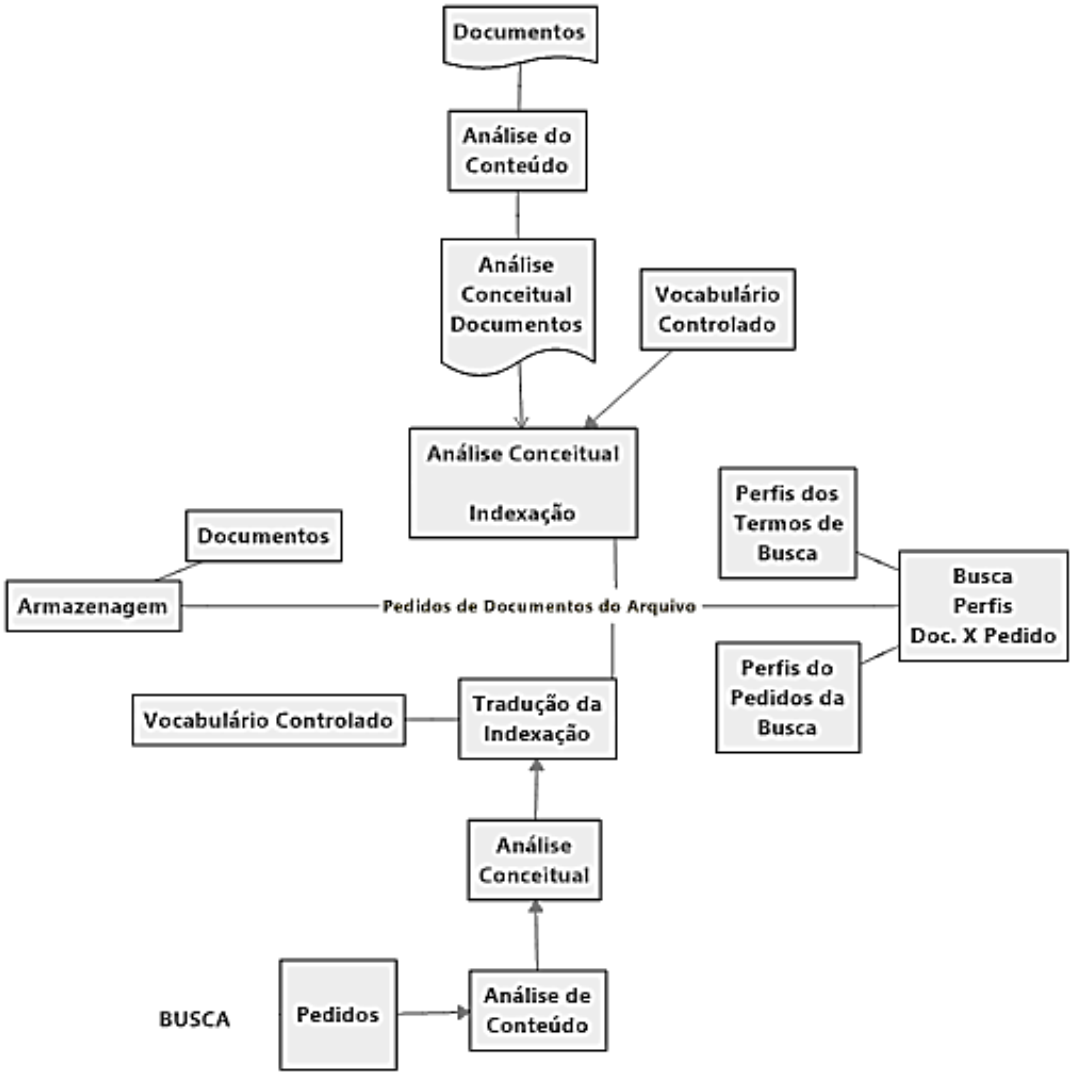
Ademais, se divide em dois tipos: classificatória (sistemas de classificações) e alfabética (cabeçalhos de assunto). Um exemplo de termo pré-coordenado seria “Televisão Digital”.

Na linguagem pós-coordenada os termos são combinados no momento da busca ou na saída, utilizando operadores booleanos como “e”, “ou” e “não” e possui algumas vantagens quando comparada a pré-coordenada, pois é mais flexível, dinâmica e permite estabelecer uma relação maior entre os termos. Entretanto, pode gerar uma falsa recuperação e exige mais conhecimento do usuário quanto a sua utilização. Um exemplo de termo pós-coordenado seria “Televisão e Digital”.

A escolha da linguagem documentária pré ou pós-coordenada deve ser pautada por “[...] vocabulários de alta especificidade, a partir de termos genéricos e específicos” e “contemplar as relações sintático semânticas entre os termos advindos das áreas científicas especializadas e da linguagem do usuário” (BOCCATO, 2009, p. 231).

A tradução não ocorre apenas no momento da indexação do documento, ela está presente também na busca realizada pelo usuário. Na busca, a tradução é feita através do indexador, ou seja, o indexador deve traduzir a necessidade do usuário para uma linguagem documentária. Para que esse tipo de tradução ocorra é necessária a interação usuário-indexador-sistema. É possível observar todo esse processo na figura 1.

Figura 1 - Processo de indexação manual



(**Fonte:** LANCASTER *apud* CESARINO; PINTO, (p. 33))

Portanto, a tradução é muito importante para a indexação. Todavia, apenas escolher a linguagem e definir os descritores que irão representar o documento não significa que tenha terminado o processo de indexação, é preciso saber se essa tradução foi feita com qualidade e isso é possível através de uma avaliação baseada em determinados critérios.

2.2.3 Normalização de termos na indexação

No processo de indexação é comum que um termo apresente diversas variações. Elas ocorrem devido a possíveis alterações com relação ao gênero, número, grau e ao possível uso de abreviações e siglas. Ademais, podem causar dispersão terminológica e descontrole no vocabulário dos documentos indexados.

Para minimizar essas variações, existem algumas ferramentas denominadas vocabulários controlados. Elas estabelecem diretrizes e princípios que auxiliam na indexação. Tais diretrizes e princípios podem ser considerados formas de normalização de termos. O vocabulário controlado é composto por uma lista de termos autorizados e constituem uma linguagem de indexação (LANCASTER, 2004, p. 19).

Segundo Cintra (2002, p. 33), as linguagens de indexação são desenvolvidas para indexação, armazenamento e recuperação da informação. Tratam-se de símbolos responsáveis por traduzir o conteúdo presente nos documentos. Portanto, os vocabulários controlados são utilizados para a tradução dos conteúdos dos documentos. Como exemplos de vocabulários controlados temos os cabeçalhos de assunto e os tesauros.

O cabeçalho de assunto é elaborado quando uma unidade informacional não desenvolveu um tesauro. Trata-se de uma lista de assuntos utilizada para criar pontos de acesso no catálogo de uma unidade informacional. Essa lista é tratada como uma convenção, pois os assuntos que a compõem foram selecionados por autoridades e são seguidas por profissionais indexadores.

O objetivo do cabeçalho de assunto é otimizar os pontos de acesso do catálogo afim de evitar a dispersão terminológica, a falta de coerência e preservar a integridade do catálogo.

O cabeçalho de assunto é composto por descritores, sinônimos e quase-sinônimos dos termos utilizados no processo de indexação. Por exemplo, “motocicleta” e “moto” possuem o mesmo significado e, portanto, são sinônimos. Se em uma unidade informacional houvesse dois documentos sobre o mesmo tema e fossem aplicados o termo “motocicleta” para um e “moto” para outro, o usuário ao realizar a busca por “motocicleta” iria recuperar apenas um documento. Logo, o cabeçalho de assunto é utilizado para que a unidade de informação escolha apenas um dos termos que irá indexar ambos os documentos. Dessa forma, o cabeçalho de assunto apresenta, para o indexador, uma regra de normalização que é a eliminação de sinônimos.

De acordo com Horner (1970 apud CESARINO; PINTO, 1978), o cabeçalho de assunto apresenta as seguintes características:

a. Linguagens estruturadas e pré-coordenadas que, de certa forma, apresentam limitações na pesquisa;

b. Os termos do vocabulário controlado são selecionados de um dicionário existente, o que o caracteriza como sistema fechado;
c. Os cabeçalhos de assuntos exercem função prescritiva
d. Linguagens não hierárquicas
e. São enumerativos, oferecendo poucas possibilidades de síntese
f. Arranjo alfabético
g. Linearidade, aplicável apenas a pesquisas unidimensionais.
h. Pouca sistemática na elaboração de cabeçalhos de assuntos e na elaboração de referências cruzadas

Segundo o acordo da Organização das Nações Unidas para a Educação (UNESCO) (1973, p. 6), o “tesauro” é um vocabulário controlado que apresenta um conjunto dinâmico de termos com relações semânticas e que abrangem um domínio específico do conhecimento. A sua função é de controle terminológico e é utilizado na tradução da linguagem natural dos documentos para uma linguagem documentária.

Grande maioria dos pesquisadores diferenciam o tesauro de um vocabulário controlado, pois determina critérios que orientam a elaboração de linguagens documentárias e diretrizes das normas (necessidade do usuário, área do conhecimento, crescimento científico, avanço tecnológico...)

De acordo com a ANSI/NISO Z39.19 (2005, p. 19), a finalidade do tesauro está relacionada a cinco objetivos:

1. Tradução: Fornecer um meio para converter a linguagem natural dos autores, indexadores e usuários para um vocabulário que pode ser usado para indexação e recuperação da informação;
2. Consistência: promover a uniformidade tanto no formato, quanto na atribuição de termos;
3. Indicação de relacionamentos: indicar relações semânticas entre os termos;
4. Rotular e pesquisar: propor hierarquias coerentes e claras em um sistema de rede, para ajudar os usuários a localizar objetos de conteúdos desejados;
5. Recuperação: servir como um auxílio na busca e localização de conteúdos informacionais.

Com relação a função desempenhada pelos tesauros, Chaumier (1989, p. 44), afirma que a sua principal função é a capacidade de representar assuntos abordados

no documento para que no momento da busca a pergunta do usuário e a resposta do sistema sejam coincidentes.

Sobre as funções desempenhadas pelos tesauros, Chaumier (1988) afirma que elas abrangem três etapas: descrição, tratamento e representação.

Quadro 5 – Etapas da função do tesauro

Descrição	Tratamento	Representação
Classifica os conceitos	Transforma a informação em dados manipuláveis.	Estabelece os conceitos da demanda.
Proporciona o vocabulário que traduz os conceitos representando-os de forma unívoca.	Favorece o controle e a validação dos dados.	Proporciona o vocabulário dos conceitos da demanda.
Favorece a coerência da análise documentária.		Facilita o diálogo entre usuário e sistema
Apresenta o entorno semântico dos descritores escolhidos para traduzir os conceitos.		

(**Fonte:** CHAUMIER, 1988)

Durante o desenvolvimento de um vocabulário controlado estão presentes diretrizes e normas que abordam os aspectos linguísticos do tesauro e que orientam na escolha e formatação dos termos que irão ser integrados ao vocabulário controlado. Algumas dessas diretrizes e normas (Quadro 6) também podem ser aplicadas para a normalização dos termos.

Quadro 6 – Diretrizes de normalização de termos

Diretriz	Função
Escolha quanto a flexão de número (singular e plural)	Tem a função de definir se os termos que irão compor o vocabulário controlado estarão no singular ou no plural.
Homógrafos (ou Polissemia)	Homógrafos são termos que possuem a mesma grafia, mas com pronúncia ou significado diferente. Por exemplo: “Serra” que pode ter o significado de instrumento cortante ou cadeia de montanhas. A função dessa diretriz é encontrar um qualificador para esses termos. Um exemplo de qualificador seria a

	utilização de um sinônimo que tenha significado único.
Sinonímia	Sinonímias são palavras diferentes, mas que possuem o mesmo significado (sinônimos). Por exemplo: “calvo” e “careca”, ambos os termos são utilizados para classificar pessoas que não possuem cabelo. A função dessa diretriz é eliminar sinônimos presentes no vocabulário controlado e assim encontrar um termo qualificador mais apropriado para a indexação.
Escolha quanto ao gênero (masculino ou feminino)	Afim de evitar a dispersão terminológica no vocabulário controlado, essa diretriz determina se o vocabulário irá utilizar apenas termos no masculino ou no feminino.
Abreviaturas e siglas	Diretriz utilizada para determinar quais abreviaturas e siglas irão compor o vocabulário controlado.
Advérbios	Em alguns casos os advérbios não devem ser utilizados como descritores dos documentos. A função dessa diretriz é eliminar frases que começam com advérbios, exceto quando possuam um significado dentro do jargão.
Verbos	Essa diretriz recomenda que verbos no infinitivo e no particípio não devem ser utilizados isoladamente como termos de indexação. Esses verbos devem ser representados por substantivos ou substantivos verbais. Exemplo: Cozinha (não “cozinhar”).

(**Fonte:** ISO 5964, 1985)

Os substantivos são os mais utilizados nos tesauros, pois possibilitam melhor descrição do termo e conceito. Em determinadas situações os substantivos podem apresentar ambiguidade, por exemplo o substantivo “prova” que pode ser tanto uma ferramenta de avaliação ou um recurso criminal.

É importante lembrar que os SNs são compostos por termos substantivados e por isso é interessante relacionar a forma como os substantivos são tratados nos tesauros e na indexação.

No tesauro os adjetivos também são utilizados como qualificadores de um termo, apresentando assim termos compostos. A principal função dos adjetivos nos tesauros é de caracterização dos substantivos.

A utilização de advérbios nos tesauros é bastante restrita, recomenda-se apenas que sejam utilizados em casos especiais, ou seja, quando agrega valor ou significado a um termo. Em algumas situações os advérbios podem ser utilizados como descritores, mas apenas quando eles possuem significado dentro do domínio em que o tesauro foi elaborado.

No tesauro é comum a presença de homonímias. Segundo Bechara (2009, p. 403), homonímias é “a propriedade de duas ou mais formas, inteiramente distintas pela significação ou função e com a mesma estrutura fonológica, os mesmos fonemas dispostos na mesma ordem e subordinados ao mesmo tipo de acentuação. Bechara (2009) afirma que dentro da homonímia é possível encontrar: homônimos homófonos e homônimos homógrafos.

Homônimos homófonos (*homo*: mesmo; fono: som) são formas que apresentam grafemas e significados diferentes, entretanto a sonoridades dos fonemas são iguais, por exemplo: sessão (espaço de tempo) e seção (local)

Homônimos homógrafos (*homo*: mesmo; grafia: escrita) apresentam a mesma grafia, mas o sentido e sonoridade são diferentes, por exemplo: força (verbo) e força (substantivo).

Existem também os Homônimos perfeitos, conhecidos também por polissêmicos, são palavras com a mesma grafia e pronúncia, por exemplo: verão (verbo) e verão (substantivo).

De acordo com a ANSI/NIZO Z39.19 (2005, p. 175) os tesauros são considerados um vocabulário controlado normalizado, portanto é importante citar a forma como os tesauros são abordados e elaborados, uma vez que o principal produto da pesquisa é a normalização de SNs para a indexação automática.

2.2.4 Avaliação da indexação

A avaliação da indexação é formada por dois métodos: avaliação intrínseca e avaliação extrínseca.

De acordo com Gil Leiva (2008, p. 385 *apud* GIL LEIVA; RUBI; FUJITA), a avaliação intrínseca é:

> [...] o conjunto de tarefas centradas no resultado da indexação (descritores, cabeçalhos, sub-cabeçalhos ou identificadores) com a finalidade de conhecer sua qualidade. A avaliação intrínseca da indexação pode ser qualitativa, isto é, por meio de valorações e consensos entre os experientes, ou quantitativa, mediante fórmulas.

A avaliação extrínseca é feita através da comparação do resultado de duas indexações, diferentes, do mesmo documento (interconsistência) e da análise da indexação com base na recuperação (exaustividade e precisão) (GIL LEIVA, 2008).

Na indexação são utilizados diversos critérios e indicadores que possibilitam avaliar sua qualidade. Alguns desses critérios são: pertinência, revocação (exaustividade), precisão (especificidade), consistência, coerência e relevância. Corroborando com essa afirmativa, Gil Leiva (2008, p. 76) apresenta o seguinte quadro com os critérios de indexação:

Quadro 7 – Critérios de indexação

Qualidades da indexação	• Exaustividade Conceitos caracterizadores do conteúdo presente no documento. • Especificidade Relação exata entre a unidade conceitual e o termo escolhido para representar. • Coerência Ausência de erros de inclusão e omissão. • Consistência Grau de coincidência entre duas ou mais indicações.

(Fonte: GIL LEIVA, 2008, p. 76, **tradução nossa**)

Os critérios de consistência e coerências estão relacionados, pois quando se busca uma indexação consistente, significa que está interessado em indexar com coerência. A coerência está relacionada aos termos atribuídos por diversos indexadores para representar o mesmo documento, quanto mais termos se coincidem, mais essa indexação está coerente. Segundo Dias e Naves (2007, p. 33), a consistência da indexação é "[...] definida como o grau de concordância na representação da informação essencial do conteúdo do documento por certos grupos

de termos de indexação, selecionados individualmente e independentemente, por cada indexador do grupo".

Para avaliar a consistência da indexação Gil Leiva (2008, p. 236), apresenta e utiliza em suas pesquisas uma variável da formula de Hooper (1965). Na formula apresentada abaixo, "T_{co}" é o número de termos comuns nas duas indexações; "A" é o número de termos na indexação A; e "B" é o número de termos na indexação B.

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

Na formula apresentada por Gil Leiva (2008), "T_{co}" é o número de termos comuns nas duas indexações; "A" é o número de termos na indexação A; e "B" é o número de termos na indexação B.

A relevância na indexação está diretamente ligada à necessidade de informação do usuário, pois é através da sua satisfação que é possível indicar o nível de relevância da indexação. Le Coadic (1996, p. 62-63) afirma que a relevância:

> "[...] mede, assim, a correspondência que existe entre um documento e uma questão. Esse conceito está na base da avaliação de desempenho dos sistemas de recuperação da informação: vincula necessidade do usuário a documento (s) e tem a ver, portanto, com a satisfação do usuário"

A pertinência se assemelha com a relevância, pois trata da relação da necessidade e do uso do documento. De acordo com Harter (1992) *apud* Dias e Naves (2007, p. 19), o critério da pertinência é "[...] outro aspecto da relevância subjetiva, tem sido usado para se referir à relação entre documento e a necessidade de seu uso".

A revocação e a precisão são usados como indicadores para a avaliação da qualidade e do desempenho, tanto do SRI como da indexação.

A revocação se trata do número de documentos relevantes para um tema que são recuperados pelo usuário no sistema. Seu valor é atribuído através do cálculo do número de documentos recuperados sobre determinado tema e o número de documentos com o mesmo tema que existem no sistema (Equação 2). Quanto mais exaustiva a indexação, maior será a revocação.

O indicador de precisão diz respeito ao número de documentos recuperados que atendem a necessidade do usuário. Seu cálculo é feito da mesma forma que a revocação, porém leva em consideração o número de documentos relevantes (Equação 3). Quanto mais específica a indexação, maior será a precisão. De acordo

com Souza (2005), os indicadores de revocação e precisão são inversamente proporcionais, ou seja, quanto maior a revocação, menor a precisão, e vice-versa (Gráfico 1).

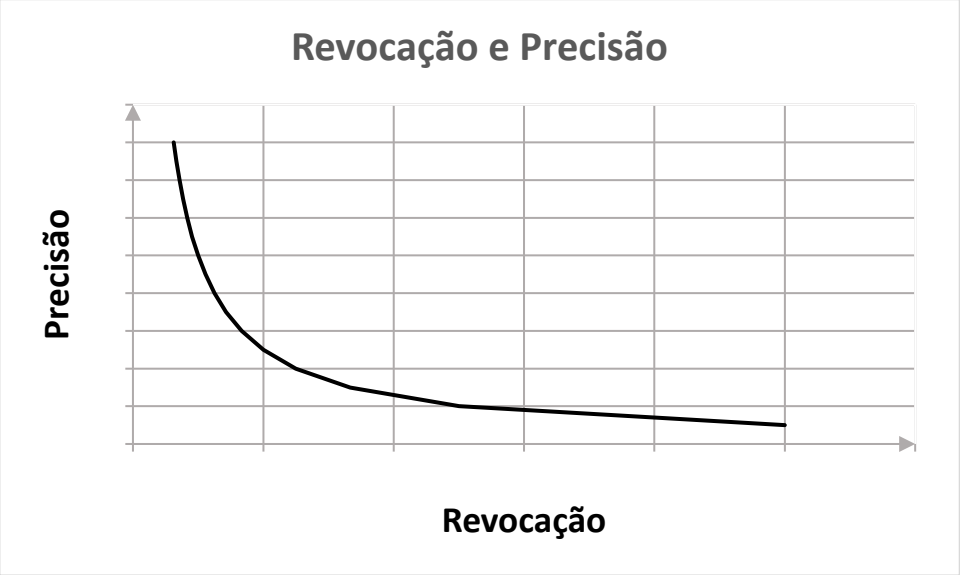
Equação 1 – Revocação

$$\text{Revocação} = \frac{\text{Nº de referências relevantes recuperadas } \times 100}{\text{Nº total de referências relevantes existentes no SRI}}$$

Equação 2 – Precisão

$$\text{Precisão} = \frac{\text{Nº de referências relevantes recuperadas } \times 100}{\text{Nº total de referências recuperadas}}$$

Gráfico 1 – Revocação e Precisão



(**Fonte:** Autoria nossa)

Além desses critérios e indicadores, Lancaster (2004, p. 36) aponta outras diretrizes para a avaliação da indexação:

> [...] O axioma da previsibilidade diz que o êxito de uma busca num sistema de recuperação depende grandemente da previsibilidade com que é descrito o conteúdo temático, o que aponta para a importância da coerência na indexação. O axioma da fidelidade diz que o outro fator que influi no desempenho é a capacidade de definir com rigor e exatidão o conteúdo temático (das necessidades de informação e, por extensão, dos documentos), que tem a ver mais com o vocabulário usado para indexar do que com a própria indexação. (LANCASTER, 2004, p. 36)

Com todas as informações apresentadas até o momento, é possível afirmar que a indexação não se trata de uma técnica, mas de uma atividade, um processo, por onde as informações apresentadas no documento são descritas para o usuário,

afim de que a informação seja recuperada. Ademais, é importante citar que a indexação não necessariamente tem que ser feita manualmente, ela pode ser feita, também, de forma automática ou semiautomática.

2.3 Indexação Automática

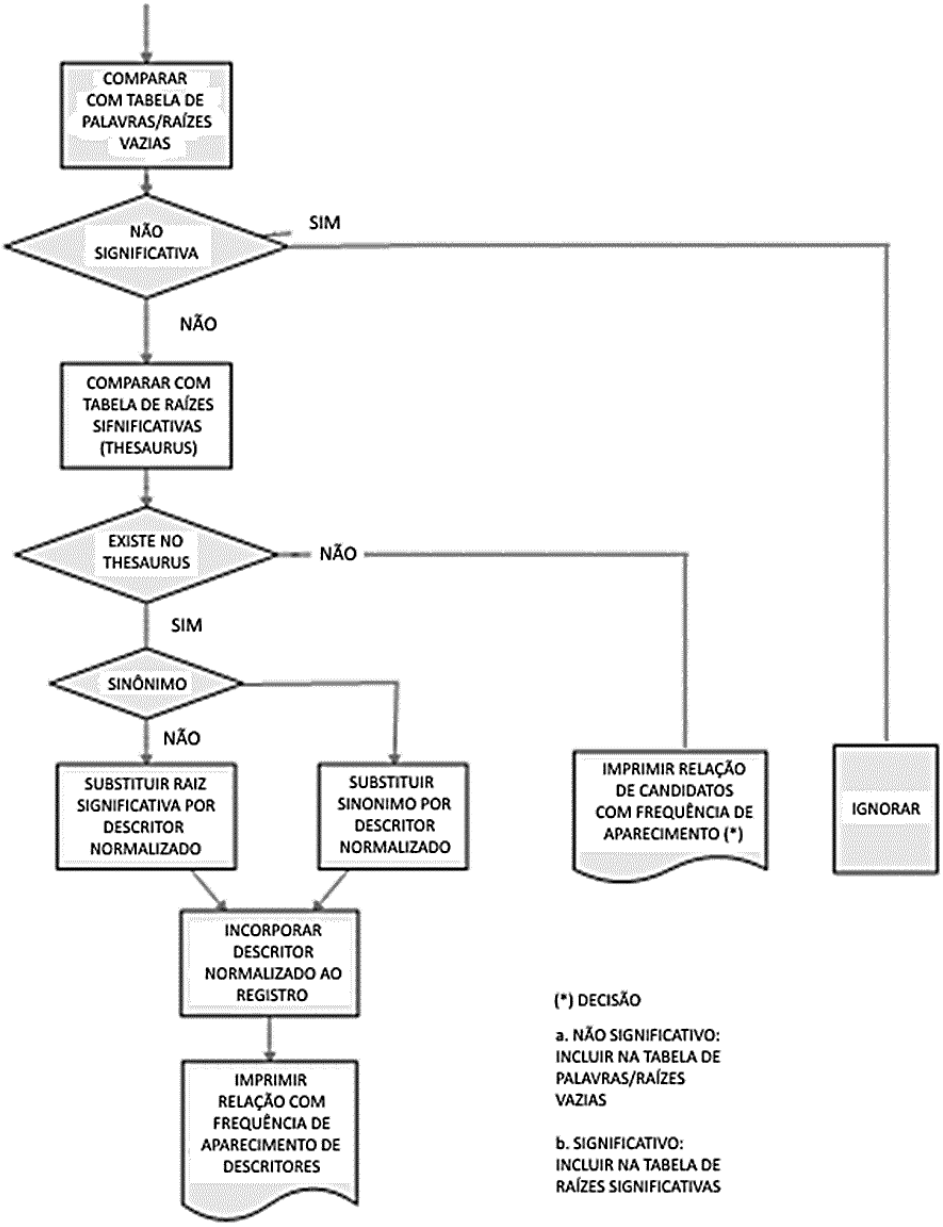
É plausível imaginar que alguns processos manuais podem ser automatizados, com a indexação isso não é diferente. Diante da grande produção informacional, pesquisadores viram que o processo de indexação manual não atenderia mais de forma satisfatória. E, a partir daí, começam a surgir os primeiros trabalhos de indexação automática.

A indexação automática tem início no final da década de 50, com o pesquisador Luhn. Ele propôs que um vocabulário presente em um documento poderia constituir uma base para a análise do conteúdo do documento. A primeira aplicação dessa proposta foi a criação do índice *Keyword in Context* (KWIC).

No Brasil, a indexação automática teve início no final da década de 60, com a utilização do programa KWIC para a elaboração de índices das bibliografias especializadas publicadas pelo Instituto Brasileiro de Bibliografia e Documentação (IBBD), conhecido atualmente como Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) (VIEIRA, 1988).

Para Robredo (1982), o processo de indexação deve comparar cada palavra presente no documento com um conjunto de palavras de significado vazio (previamente selecionadas), e isso irá ajudar o processo de eliminação, já que o programa de computador irá considerar as palavras restantes como palavras significativas (Figura 2).

Figura 2 — Processo de indexação automática



(**Fonte:** ROBREDO, 1982, p. 247)

A indexação automática (também conhecida como indexação assistida por computador, ou por indexação semiautomática) é considerada um modelo de extração com características estatísticas e probabilísticas (BORGES, 2009).

Narukawa, Gil Leiva e Fujita (2009), apresentam três conceitos diferentes que estão presentes na indexação:

1. **Indexação assistida por computador** – São utilizados programas de computador que armazenam os termos extraídos pelo profissional.

2. **Indexação semiautomática** – Programas de computador são responsáveis pela a análise dos documentos (automaticamente) e quando necessários os termos são avaliados por um profissional.
3. **Indexação automática** – Os programas realizam o processo de análise dos documentos sem a validação dos profissionais.

Diversos autores definiram o que seria a indexação automática. De acordo com Vieira (1988), a indexação automática é uma tarefa que envolve um computador que é responsável por analisar os textos e construir um índice de assuntos, possibilitando a sua recuperação. Hjørland (2008) define a indexação automática como um procedimento realizado por algoritmos que funcionam em uma base de dados onde estão presentes as representações dos documentos (textos completos ou parciais, registros bibliográficos, etc.).

Percebe-se que a indexação automática tem como principal característica o auxílio do computador, e isso foi pensado com o objetivo de agilizar a indexação manual. Porém, assim como a indexação manual, a indexação automática também é passível de problemas, em sua grande maioria causados pelo próprio sistema que, dependendo da sua configuração de indexação, faz com que o documento não possa ser recuperado (GIL LEIVA; FUJITA, 2012).

Lancaster (2004, p. 56-59) cita um sistema importante para a indexação automática: o sistema *Nested Phrase Indexing System* (NEPHIS). O NEPHIS foi criado em 1977 e apresentava um índice articulado de assuntos, criado por Timothy C. Craven. Nesse sistema, os termos eram reordenados para que se ligassem ao seu vizinho original, através de palavras funcionais ou pontuações, apresentando assim uma estrutura similar a frase.

Nessa época já era evidente a relação entre a indexação e os aspectos linguísticos (SALTON, 1973). Alguns pesquisadores já afirmavam que mais trabalhos deveriam focar os estudos das propriedades estruturais e semânticas das línguas naturais.

Na literatura são apontados diversos outros tipos de indexação automática, entre eles estão: a indexação por extração automática, a indexação por atribuição automática.

O processo de indexação por extração automática baseia-se na frequência e na posição em que certas palavras ou expressões aparecem no documento, sendo então extraídas para representá-lo (LANCASTER, 2004).

A indexação automática por atribuição é um processo mais complicado, mas que apresenta uma eficiência maior quando comparada ao processo de extração (BORGES; MACULAN; LIMA, 2008). No processo de indexação automática por atribuição é necessária a utilização de um instrumento de controle terminológico (vocabulários controlados). A grande dificuldade encontrada nesse processo de indexação é que para os programas de computador é muito difícil uma atribuição correta dos termos. Por exemplo, a frase “Após a queda do muro, a Alemanha ocidental e a Alemanha oriental...” na indexação manual pode ser indexada como Guerra Fria, mas para o computador esse nível de interpretação é mais difícil (O’CONNOR, 1965 *apud* LANCASTER, 2004).

Autores como Edmundson (1969), Garvin (1969) e Salton (1973), percebiam uma relação entre o processamento da informação e a linguística computacional, e defendiam que os estudos desenvolvidos nessa área deveriam ser voltados as propriedades estruturais e semânticas da linguagem natural. Percebia-se que as relações semânticas eram importantes, pois tinham uma estruturação do conhecimento e uma formação de conceitos que possibilitavam a escolha de termos representativos que possuíam significado.

Interessados na questão da semântica e da sintaxe no processo de indexação automática, autores como Kuramoto (1995), Souza (2005), Borges (2009), Silva (2014), Silva e Corrêa (2015) e Nascimento (2015), destacam a importância da semântica e da sintaxe para a indexação automática afim de melhorá-la, se preocupando quanto ao significado das palavras atribuídas aos documentos.

A importância da indexação automática, com relação a semântica e a sintaxe, se deve a possibilidade do *software* identificar as estruturas léxicas das frases, e os significados dos termos que descrevem o documento. Com a sintaxe é possível determinar a forma correta de estruturação das frases, através da sequência de sujeitos, verbos, objetos, predicados, etc. Já a semântica é responsável por atribuir significado a frase construída. Segundo Sautchuk (2010), a sintaxe

> “[...] 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).

Assim como a indexação manual, a indexação automática utiliza os termos retirados de um documento para construir um índice que possa descrevê-lo. Esses

termos retirados automaticamente são considerados como termos isolados, o que diminui o valor informacional. De acordo com Maia (2008), a indexação tem objetivo de representar o conteúdo de um documento, através de uma lista de termos (descritores). Logo, "[...] os descritores devem, na maior extensão possível, ser portadores de informação, de maneira a relacionar um objeto da realidade extralinguística com o documento que traz informações sobre esse objeto" (MAIA, 2008, p. 28).

Kuramoto (2002) afirma que a maioria dos SRIs utilizam palavras isoladas para descrever a informação e isso traz alguns problemas linguísticos, como: a polissemia[4], a sinonímia[5], e as palavras combinadas que, em ordem diferente, transmitem diferentes significados. Com base nesses problemas, o autor concluiu que na busca em um SRI, a polissemia e a combinação de palavras aumentavam a taxa de ruído[6], e a sinonímia aumentava a taxa de silêncio[7].

A partir das informações dispostas, como a importância da semântica e da sintaxe, e dos problemas ocasionados pelas palavras isoladas, Kuramoto (2002) aponta como solução desses problemas a utilização dos SNs como descritores do conteúdo. O sintagma nominal "é a menor parte do discurso portadora de informação" (KURAMOTO, 1995) e quando extraído do texto mantém o significado e o conceito. Perini (1998) corrobora ao afirmar que os SNs eram mais eficientes na recuperação da informação e na classificação se comparados as palavras isoladas.

2.4 Sintagmas Nominais

Para entendermos SNs é necessário compreendermos o que são os sintagmas.

Sintagmas são unidades de significados atribuídos a orações e que as organizam de acordo com leis sintagmáticas. Perini (1998) entende sintagmas como grupos de unidades que compõem sequências maiores, mostrando coesão entre eles.

Segundo Othero (2009) uma estrutura sintática segue determinadas regras, não é apenas um conjunto disperso de palavras. As posições de cada palavra são importantes, pois lhe dão sentido, e a organização das palavras dentro de uma sentença pode formar um sintagma.

[4] A palavra possui diversos significados. Exemplo: Reflexo (movimento ou imagem espelhada?).
[5] Palavras diferentes com o mesmo significado. Exemplo: Macaxeira e Mandioca.
[6] Está relacionado a taxa de precisão.
[7] Está relacionado a taxa de revocação.

Perini (2005, p. 44-45) apresenta o seguinte exemplo de sintagma: “A casa de Lulu é azul e branca”. De acordo com o autor, “A casa de Lulu” forma uma unidade, pois há uma coesão, ao passo que “Lulu é azul” não forma, pois não possui coesão. Ainda segundo o autor, as frases são formadas por constituintes, e a eles são atribuídas funções na análise tradicional, ou seja, “A casa de Lulu” é uma constituinte, enquanto “Lulu é azul” não. Por último, o autor demonstra os possíveis constituintes do sintagma: “casa de Lulu”, “azul e branca”, “é azul e branca”, etc.

Os sintagmas são subdivisões naturais que ocorrem nas orações. No âmbito semântico, os sintagmas se tratam de uma unidade que possui significado único e coerente, e são classificados de acordo com as funções que ocupam. Quando lhe é atribuída a função de substantivos (sujeito, objeto), são classificados como SNs. Por outro lado, se desempenham função de predicado, são classificados como sintagmas verbais (SV) (PERINI, 2005, p. 43-44).

Sintagmas definem relações de dependência e estabelecem ordens de subordinação para cada elemento presente na frase. O termo sintagma é empregado para designar partes de uma oração. Os tipos de sintagma são:

- Sintagma nominal;
- Sintagma adjetival;
- Sintagma verbal;
- Sintagma preposicional;
- Sintagma adverbial.

Segundo Perini (1998), os SNs são compostos por duas estruturas. Essas estruturas são identificadas com base no núcleo e por isso são definidas como estruturas pré ou pós-nucleares. Uma estrutura pré-nuclear pode ser composta pelos seguintes elementos: predeterminantes, determinantes, quantificadores, possessivos sintéticos, e numeral. Já a estrutura pós-nuclear pode ser composta por palavras que qualifiquem o núcleo (modificadores).

Quadro 8 – Elementos dos sintagmas nominais

Elementos que compõem os SNs		
Elementos pré-nucleares	**Núcleo**	**Elementos pós-nucleares**
Predeterminantes, determinantes, quantificadores, possessivos sintéticos, numeral.	Nome (substantivo, pronome substantivo, numeral ou palavra substantivada)	Modificadores (palavra ou conjunto de palavras que qualificam o núcleo, restringem o sentido do núcleo, inclusive outros nomes que podem ser núcleos também)

(**Fonte:** PERINI, 2010, p. 259).

A estrutura de um sintagma nominal é formada, obrigatoriamente, por apenas um núcleo e, em alguns casos, é acompanhada por determinantes ou adjacentes (elementos opcionais) (Quadro 9).

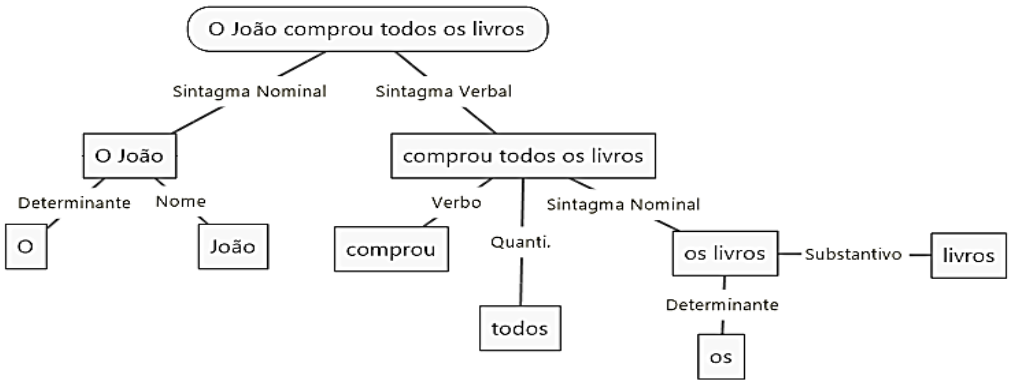
Quadro 9 – Estrutura do sintagma nominal

Estrutura – Sintagma Nominal		
Determinantes (opcional)	**Núcleo (obrigatório)**	**Adjacentes (opcional)**
Artigos	Substantivo	Adjetivo
Indefinidos	Pronomes	Nomes
Demonstrativos	Palavras Substantivadas	Advérbios
Numerais	Preposições Substantivadas	Preposições Adjetivais
Interrogativos		Construções Preposicionais
Exclamativos		
Relativos		

(**Fonte:** desenvolvido pelo autor.)

Para melhor compreender um sintagma nominal, observe o exemplo diagramado na figura 3.

Figura 3 – Exemplo de sintagma nominal



(**Fonte:** Autoria nossa)

Analisando o exemplo da figura 3, percebe-se que em uma mesma oração os SNs podem aparecer diversas vezes, ora como integrantes de outros sintagmas, ora ligados por uma preposição (SNs preposicionados). No caso exemplificado, o SN está inserido em um sintagma verbal.

As estruturas sintáticas são formadas por estruturas básicas que possibilitam infinitas combinações. Apesar disso, é importante ressaltar que os SNs possuem diversas regras quanto à sua formação. Fundamentado por Miorelli (2001) e Santos (2005), Silva (2014, p. 50) elaborou um conjunto de regras quanto à formação dos SNs:

Figura 4 – Regras de formação dos sintagmas nominais

Regras	Exemplos
Regra geral: DET + MOD + N + MOD	A interdisciplinar Ciência da Informação
Regra 1: DET + N + MOD	A Ciência da Informação
Regra 2: N + MOD	Informação estratégica
Regra 4: DET + N	A informação
Regra 5: N	Informação
Regra 6: DET + N + DET + N + MOD	A filosofia e a ciência juntas
Regra 7: DET + DET + N + MOD	A minha recuperação da informação
Regra 8: MOD + N + MOD	Grande área da informação
Regra 9: DET + DET + N	Uma certa área

(**Fonte:** Silva (2014, p. 50), baseado em Miorelli (2001) e Santos (2005).)

De acordo com Kuramoto (1995), os SNs podem ser compostos por outros SNs, sendo então classificados em níveis. Seus níveis são definidos de acordo com a quantidade de SNs que apresentam. Por exemplo, um SN que não possua outro SN em sua estrutura, é considerado nível 1. Já um SN que possua outro SN de nível 1 em sua estrutura, é considerado nível 2, e assim sucessivamente. Na prática,

teríamos: “Sistema de Classificação de Documentos” (SN nível 3), “Classificação de Documentos” (SN nível 2) e “Documentos” (SN nível 1).

De acordo com Perini (1996), sintagma nominal faz parte de uma classe gramatical que possui o comportamento de sujeito da oração, de objeto direto e, se vier após um predicado, de adjunto adnominal ou de objeto indireto.

No caso do sintagma verbal, o seu núcleo é um verbo; sintagma adjetival, o núcleo é um adjetivo, e os sintagmas preposicionados, são constituídos de uma preposição acompanhado de um sintagma nominal

2.4.1 Indexação Automática por Sintagmas Nominais

A análise dos aspectos sintáticos e semânticos do documento de forma automática, através de um computador, torna-se um grande diferencial na indexação.

No processo clássico de indexação automática, percebia-se que o uso de palavras isoladas vinha apresentando resultados insatisfatórios quando se tratava de representação e recuperação da informação. Com base nesse problema, começou a se estudar os aspectos semânticos e sintáticos para a indexação automática, a fim de tornar o processo de recuperação da informação mais eficiente. Como resultado de alguns desses estudos, foi apresentada a utilização de grupos de substantivos (*noun groups)*, uma vez que substantivos possuem maior valor semântico se comparado a outros termos morfológicos da linguagem. Em decorrência dos estudos do uso dos substantivos, surgem os estudos sobre os sintagmas nominais (*noun phrases*) e começou a se perceber que a sua aplicação no processo de recuperação e representação da informação obtinha resultados satisfatórios.

Os pesquisadores que se destacam quanto a indexação automática por SNs para textos em língua portuguesa são: Kuramoto (1995, 1996, 2002), Miorelli (2001), Souza (2005), Santos (2005), Corrêa *et al* (2011), Silva (2014), Martins (2014) e Nascimento (2015). Porém, o ponto de partida dessa temática (SNs) foi dado por Michel Le Guern (1991). Sua proposta era trocar as palavras por SNs como descritores da informação pois, diferente das palavras, os SNs eram portadores de significado para a indexação e recuperação da informação. Ele ressaltava a diferença que existia entre o descritor e a palavra, sendo o descritor uma unidade do discurso e palavra apenas uma unidade da língua, já que não possuí significado para a indexação e recuperação da informação.

Com base nisso, a diferença da indexação automática com base em palavras e a baseada em SNs é quanto a sua significação. O processo de indexação, seja ele automático ou não, deve extrair do documento descritores que facilitem sua recuperação, e não símbolos sem referências, como são as palavras (KURAMOTO, 2002). Segundo Kuramoto (2002, p.6), os SNs são a menor unidade do discurso portadora de informação e podem se apresentar como uma palavra isolada ou um conjunto de palavras, ambos com valor semântico e sintático.

Portanto, para que o SRI tenha melhor desempenho é necessário que a indexação seja feita não apenas com base em palavras sem referência, mas que o descritor do documento seja uma unidade do discurso. Ou seja, ao invés de palavras, os sistemas de indexação automática deveriam extrair e utilizar unidades do discurso.

Miorelli (2001) aponta que os conjuntos de frases presentes nos documentos podem ser utilizadas para a construção de índices, e para isso é preciso selecionar cada conjunto. Dessa forma, as palavras isoladas consideradas sem referência podem ser substituídas pelos SNs, com o objetivo de construção de um índice hierárquico que pode se transformar em uma interface de busca para o usuário, ideia essa proposta por Kuramoto (2002):

> A primeira seria implementar uma indexação automática nos moldes daquela tradicional baseada em palavras, substituindo os índices das palavras isoladas por índices dos sintagmas nominais... Uma segunda alternativa seria o aproveitamento da organização hierárquica, em árvore, dos sintagmas nominais.

A partir desses questionamentos quanto a substituição de palavras isoladas por SNs para realizar a indexação automática, foram desenvolvidas diversas pesquisas para a identificação e extração dos SNs através de programas de computador (*softwares*). A extração e identificação de SNs por um ser humano pode ser considerada uma atividade simples, porém para a máquina pode ser bastante complexa, pois é necessário descobrir formas de “ensiná-la” a reconhecer os SNs.

Um dos pioneiros na pesquisa de SNs para a indexação automática para textos em português foi Kuramoto que, em sua pesquisa, em 1995, aplicou técnicas para a extração e identificação dos SNs. Contudo, essa pesquisa foi direcionada para a extração e identificação de forma manual, ou seja, realizada pelo homem, mas é muito importante citá-la, pois para que uma máquina consiga reconhecer os SNs é necessário que primeiro tenhamos conhecimentos das dificuldades existentes na

realização dessa atividade. As dificuldades destacadas por Kuramoto (1995), no processo de extração dos SNs foram:

1. **Sintagmas nominais escondidos em frases com fatoração –** Frases com fatoração são as que possuem palavras procedidas por outra sequência de palavras ligadas por conjunções (“e”, “ou”, etc.). A dificuldade está relacionada a identificar se o SN é único ou se trata de um conjunto de SNs. Por exemplo, na frase “As dívidas das empresas públicas e privadas” são identificados os seguintes SNs: “As dívidas das empresas públicas” e “As dívidas das empresas privadas”. Outro tipo de frases com fatoração são as procedidas por parênteses. Por exemplo, a frase “Conjunto de Leis (penais, trabalhistas, comerciais e internacionais)” possui os seguintes sintagmas: “Conjunto de leis penais”, “Conjunto de leis trabalhistas”, “Conjunto de leis comerciais” e “Conjunto de leis internacionais”.
2. **Artigo zero –** Está relacionado aos determinantes. Na língua portuguesa é comum a falta do determinante no SN e quando isso ocorre ele é classificado como SN de artigo zero.
3. **Cálculo de anáforas –** Segundo Gasperin, Goulart e Vieira (2003), no texto, quando uma expressão é utilizada da segunda vez em diante, para referenciar outra, ela é considerada uma anáfora. Isso posto, os principais problemas foram com relação as anáforas sem fonte explícita e as anáforas com fonte no tempo ou nos acontecimentos. Exemplo de anáfora sem fonte explícita: “baseado nisso” (nisso o quê?). Exemplo de anáforas com fonte no tempo ou nos acontecimentos: “nesse período” (que período?), “esses sistemas de extração” (esses quais?).
4. **Cálculo de elipses –** Trata-se da capacidade de perceber a falta de determinadas palavras em um contexto de uma oração, por exemplo: “abordar o usuário é importante não só para avaliação (?), mas também para evolução do sistema”. Percebe-se que é necessário o complemento (do sistema) para identificar qual o SN.

Como dito anteriormente, é importante destacar as dificuldades encontradas na extração de SNs pelo ser humano para que sejam desenvolvidos sistemas de extração automática com melhor qualidade. Logo, é importante que o sistema seja alimentado com os conhecimentos (morfológicos, sintáticos e semânticos) obtidos

através de pesquisas desenvolvidas pelo ser humano, possibilitando uma melhor análise dos textos com o objetivo de, automaticamente, selecionar e extrair os SNs, de forma correta.

Diversos pesquisadores se destacaram com seus estudos no âmbito da extração e seleção automática dos SNs. Além de Kuramoto (1995), que desenvolveu um protótipo de um SRI baseado na navegação de SN estruturados, outros pesquisadores também obtiveram destaque por suas pesquisas, tais como: Bick (2000) que se destaca pela criação do *software* PALAVRAS; Souza (2005), Borges, Maculan e Lima (2008), Maia e Souza (2010), Corrêa et. al. (2011), Silva (2014) e Nascimento (2015), por estudos sobre a indexação automática por SNs; e Vieira et. al. (2000), Miorelli (2001), Lopes (2012), Souza e Raghavan (2014), por estudos sobre a extração e seleção de SNs.

Em algumas dessas pesquisas foram utilizadas algumas ferramentas que auxiliaram no processo de indexação automática baseada em SNs, como: Etiquetadores, Identificadores de SNs, Extratores de SNs e Selecionadores de SNs. Com base nessas ferramentas e em pesquisas que avaliavam a utilização delas, foi possível desenvolver *softwares* específicos para a realização dessas atividades. Nascimento (2015) apresenta o seguinte quadro descrevendo a funções dessas ferramentas.

Quadro 10 – Ferramentas da indexação automática

Ferramentas/sistemas	**Funções desempenhadas**
Etiquetadores **1ª Ferramenta: Etiquetagem das palavras do texto.**	Os etiquetadores (Taggers) têm como função identificar e rotular as palavras que compõem um texto em determinadas classes gramaticais. Categoriza as palavras e as rotulam através de etiquetas, para que essas palavras etiquetadas passem para a ferramenta seguinte, que é a identificação dos SNs. Como exemplo de etiquetadores, tem-se o FORMA que é um programa de código aberto. Outro software bem utilizado é o parser PALAVRAS, entre outros.
Identificadores de SNs **2ª Ferramenta: Identificação dos SNs.**	Programas computacionais que têm como propósito analisar um determinado texto, observando as sequências de léxicos e aplicando esses léxicos às regras gramaticais internas dos SNs (gramática sintagmática) com o propósito de que sejam identificadas as sequências de palavras que constituam SNs. Exemplo: PALAVRAS
Extratores de SNs **3ª Ferramenta: Extração de SNs.**	Os extratores são ferramentas desenvolvidas para realizar, além da atividade de identificação dos SNs, a extração dos mesmos, mostrando os fora do texto. Exemplo: OGMA

Selecionadores de SNs 4ª Ferramenta: Seleção dos SNS.	A seleção se faz necessária, pois, muitas vezes, os SNs extraídos pela máquina não são representativos do conteúdo de determinado documento. Em suma, essa função escolhe, com base em determinados critérios, SNs com valor de descritores dentre os vários SNs extraídos pela máquina, ou seja, os melhores SNs.

(**Fonte**: Nascimento, 2015)

Verifica-se na literatura de indexação automática baseada em SNs, que as pesquisas desenvolvidas focavam exclusivamente na identificação, extração e seleção dos SNs para a sua aplicação na indexação automática, porém é importante questionar se todos os SNs extraídos e selecionados são semelhantes a descritores documentais? Ou podem ser traduzidos em descritores documentais? Sendo os descritores documentais os termos de indexação retirados de uma linguagem de indexação.

Com base nas pesquisas de Corrêa *et al*. (2011), Silva (2014) e Nascimento (2015), é possível afirmar que nem todos os SNs extraídos dos documentos possuem valor semântico e sintático para serem definidos como descritores do conteúdo presente no documento. Com base nisso, no quadro 9 o processo de indexação automática por meio de SNs, apresentado por Nascimento (2015), é modificado com a adição de mais uma etapa a esse processo, e ela se refere a normalização dos SNs em termos canonizados. Portanto, o novo quadro de processos de indexação automática por meio de SNs ficaria da seguinte forma:

Quadro 11 – Etapas da indexação automática com base em sintagmas nominais

Processo de indexação automática por meio de SNs	
1ª Etapa	Identificação dos SNs através das subetapas de "etiquetagem" e de "cotejamento dos léxicos etiquetados com as regras dos SNs"
2ª Etapa	Extração dos SNs do texto, mostrando-os em listas, por exemplo.
3ª Etapa	Normalização dos SNs em termos canonizados
4ª Etapa	Seleção dos SNs, normalizados com base em critérios que os classifiquem como "Bons Descritores"

(**Fonte:** desenvolvido pelo autor, baseado em Nascimento,2015).

2.4.2 Extração e Seleção dos Sintagmas Nominais

Essa seção tem como foco mostrar algumas pesquisas que se destacaram no âmbito da indexação automática com SNs e assim apresentar os critérios, princípios,

heurísticas ou métodos utilizados para a extração e seleção dos SNs para avaliar a possibilidade de aplicá-los como critérios para a normalização dos SNs.

Atualmente, é possível fazer a extração dos SNs através de programas de computador como o OGMA e PyPLN. Mas a extração não é o suficiente, é necessário também um processo de normalização desses sintagmas para que sejam transformados em termos e assim utilizados na indexação automática.

Um fator importante para a recuperação da informação é a escolha dos termos extraídos do documento, essa escolha deve ter um valor conceitual e não apenas terminológico. Ao considerar esse valor conceitual, temos inserido o papel dos SNs. Todo sintagma é portador de informações conceituais, ou seja, todo sintagma nominal é, a princípio, um termo candidato a conceito (LOPES, 2012).

Segundo Lopes (2012, p.33), “Um aspecto importante para a recuperação de informações textuais é o passo posterior à extração de termos, que consiste em escolher dentre os termos extraídos aqueles que são portadores de valor conceitual, e não apenas terminológico”.

Na tese desenvolvida por Souza (2005) foi proposta uma metodologia para a indexação através da identificação, extração e seleção de SNs dos textos completos utilizados em documentos digitais. Foram utilizados dois *corpora* de documentos para experiência da metodologia. O primeiro *corpus*, que também foi utilizado por Kuramoto (1999), possui 15 documentos e foi utilizado para validar a extração automática dos SNs. O segundo corpus, composto por 60 documentos, foi utilizado para validação da metodologia proposta por Souza (2005). Para validar a metodologia, Souza (2005) dividiu o corpus em dois igualmente proporcionais, onde o primeiro seria utilizado para a metodologia prospectiva e o segundo para a metodologia consolidada.

Após a aplicação da metodologia prospectiva, Souza (2005) analisou os dados e efetuou algumas mudanças. Com base nessas mudanças, Souza (2005) desenvolveu a metodologia consolidada, aplicou-a e, por fim, avaliou e comparou ambas as metodologias.

Na pesquisa desenvolvida por Souza (2005) foram apresentados alguns critérios para a seleção dos SNs que podem auxiliar no seu processo de normalização. Os critérios que mais se destacaram foram:

1. Frequência de ocorrência dos SNs nos documentos;
2. Descarte de SNs com frequência inferior a um patamar preestabelecido;

3. Análise manual de SNs pré-escolhidos (decisão de relevância);
4. Análise da estrutura e nível do SNs;
5. Verificação do SNs em tesauro específico.

Outra pesquisa que se destaca na extração de SNs é a de Maia (2008). Na sua pesquisa foi proposto o desenvolvimento da ferramenta “OGMA”, que tinha como função realizar automaticamente a extração de SNs e calcular a pontuação de cada SN na indexação de documentos. Sua tese tinha como objetivo investigar a utilização de SNs “pontuados” para a classificação automática por similaridade e aglomerados de documentos eletrônicos.

A pesquisa de Maia (2008) se divide em três etapas: a primeira é voltada ao desenvolvimento da ferramenta OGMA, a segunda tem como objetivo a aplicação da metodologia prospectiva e a terceira é a aplicação da metodologia consolidada.

Maia (2008) utilizou em sua pesquisa dois corpora, sendo o primeiro composto por 50 artigos do Encontro Nacional de Pesquisa em Ciência da Informação, e o segundo por textos menores, de cunho jornalístico e de outros temas. Assim como Souza (2005), Maia (2008) utilizou da metodologia prospectiva e da metodologia consolidada em sua pesquisa. A metodologia prospectiva foi aplicada para o primeiro corpus e a metodologia consolidada foi aplicada para o segundo corpus. O autor explica que a utilização de corpus diferentes foi para avaliar e perceber como cada corpus se comportava.

Com relação ao processo de seleção de SNs que possuem valor de descritores, Maia (2008) pontuou os SNs levando em conta quatro critérios: frequência dos SNs no documento; a incidência dos SN em outros documentos; os níveis dos SNs e as estruturas sintáticas dos SNs.

Na pesquisa de Corrêa *et al*. (2011), verificou-se como os SNs podem ser utilizados no processo de indexação e recuperação de informações em meio digital. Foi realizado um estudo de caso no âmbito da Biblioteca Digital de Teses e Dissertações da UFPE, com o objetivo de analisar a extração de SNs e sua utilização como ponto de acesso à informação.

Para o desenvolvimento dessa pesquisa foram utilizados como corpus 30 resumos (divididos igualmente) de três programas de pós-graduação (Direito, Ciência da Computação e Nutrição). A ferramenta utilizada para a extração dos SNs foi a desenvolvida pela pesquisa de Maia (2008), o OGMA.

Na pesquisa foi avaliado o processo de extração dos SNs realizado através da ferramenta OGMA. A avaliação foi feita com base no cálculo e na análise dos percentuais de precisão dos SNs relevantes como descritores extraídos; da taxa de erro na extração de caracteres que não caracterizam SNs e no percentual de SNs extraídos, mas que não são relevantes como descritores.

Ao término da pesquisa, Corrêa *et al*. (2011) afirmam que deve ser dada mais atenção para a ordenação por relevância dos SNs. Segundo os autores, a extração dos SNs não é por si só suficiente para a indexação e recuperação de informação em ambientes digitais, pois alguns SNs extraídos não possuem valor representativo para o documento. Isso também é constatado por Kuramoto (2002) e Lopes (2012), quando discutida a importância da seleção de SNs extraídos.

Na pesquisa feita por Souza e Raghavan (2014), é abordada a seleção de SNs com base na semântica intrínseca do texto. Os principais objetivos dessa pesquisa foram: 1) avaliar a utilidade e o valor de uma metodologia desenvolvida para extrair e atribuir pontuação (pesos) aos SNs em língua portuguesa; 2) examinar o valor e a utilidade dos SNs mais bem classificados, que podem servir como descritores ou palavras-chave para representar o "*aboutness*" dos documentos os quais os SNs foram extraídos.

Os autores consideraram três (3) critérios para o cálculo do *score* (pontuação) de cada sintagma nominal, que foram: a frequência do SN no documento, a frequência inversa do SN nos documentos e o valor do SN de acordo com a CNP (*Category of Noun Phrase* – Categoria do Sintagma Nominal). O CNP se baseia no nível e na estrutura do sintagma nominal.

O *corpus* utilizado por Souza e Raghavan (2014) foi o mesmo utilizado em suas pesquisas anteriores (Souza, 2005; Souza; Raghavan, 2006), e abrange o total de 60 artigos científicos oriundos de revistas da área de CI. A ferramenta utilizada para o processamento dos textos foi o PALAVRAS e o PALAVRAS xtractor.

Durante a pesquisa, foram utilizados quatro (4) métodos diferentes para o cálculo e a atribuição do score de cada sintagma nominal extraído, levando em consideração: a frequência normalizada; a frequência normalizada e a frequência inversa; a frequência normalizada a frequência inversa e o fator CNP; e a função OKAPI BM25.

No âmbito da recuperação da informação, a OKAPI BM25 se trata de uma função de classificação (*ranking*) utilizada por motores de busca, com o objetivo de

classificar documentos de acordo com sua relevância. Essa função é baseada em uma estrutura de recuperação probabilística desenvolvida nas décadas de 70 e 80 por diversos cientistas, entre eles Stephen E. Robertson e Karen Spärck Jones.

A frequência normalizada, apresentada por Souza e Raghavan (2014), tem o propósito de corrigir distorções relacionadas ao comprimento do documento e é calculada através da razão entre a frequência absoluta do SN no documento e o número total de SNs presentes no documento.

Após a aplicação de cada método, Souza e Raghavan (2014) fizeram uma análise manual dos SNs que foram selecionados, com objetivo de verificar o valor descritivo de cada SN. Para verificar esse valor, cada SN foi comparado com as palavras-chave atribuídas ao documento pelo autor original. Com a conclusão da análise, os SNs foram classificados em: altamente relevantes; razoavelmente relevantes; moderadamente relevantes; e não relevantes.

O resultado obtido com a pesquisa foi que a atribuição de pesos aos SNs com base apenas na frequência não é satisfatória. Todavia, percebeu-se uma melhora ao se utilizar a frequência normalizada e a frequência inversa no documento para o cálculo de pesos, e uma melhora um pouco maior ao utilizá-las em conjunto com a CNP. Com os pesos atribuídos com a OKAPI, foi percebido um aumento no número de SNs selecionados, pois muitos SNs tiveram o mesmo peso.

Quando comparado os resultados dessa pesquisa (SOUZA e RAGHAVAN, 2014) com os resultados da pesquisa anterior (SOUZA e RAGHAVAN, 2006), é observado que os resultados da pesquisa desenvolvida em 2006 foram mais satisfatórios, podendo esses resultados estarem relacionados a presença de uma lista de *stopwords* utilizada para descartar os SNs menos relevantes. Os autores concluíram e indicaram, para pesquisas futuras, a utilização de uma lista de *stopwords*, com a finalidade de suprimir a geração de pontos de acesso e a recuperação de SNs de valor ou utilidade questionáveis.

Souza e Raghavan (2014) observaram uma melhora nos resultados obtidos com da utilização da CNP, e com isso decidiram testar diversos valores de CNP afim de encontrar seu valor ideal. Através desses testes, verificou-se resultados mais satisfatórios com os seguintes valores de CNP:

Quadro 12 – Valores otimizados – CNP

Categoria	Estrutura e nível do SN	Valor CNP
1a	Nível 1, estrutura (D*+N)	0,2
1b	Nível 1, qualquer estrutura, exceto (D*+N)	0,8
2	Nível 2, qualquer estrutura	1,1
3	Nível 3, qualquer estrutura	1,4
4	Nível 4, qualquer estrutura	1,2
>4	Nível 5, ou superior a qualquer estrutura	0,8

(**Fonte**: Souza e Raghavan (2014, p. 14, **Tradução nossa**))

Na pesquisa de Silva (2014) foi feito um levantamento do estado da arte da indexação automática baseada em SNs em textos de língua portuguesa. Os objetivos desse levantamento são: 1) discutir os critérios de seleção e extração de SNs como descritores documentais; 2) avaliar e comparar as ferramentas de extração automática de SNs: Parser, PALAVRAS, OGMA e LX-Parser, utilizando como referência a extração manual de SNs, com o objetivo de destacar ferramentas e métodos que possam auxiliar na indexação automática.

Para o desenvolvimento da pesquisa, Silva (2014) teve que realizar, primeiramente, um levantamento bibliográfico do estado da arte da indexação automática baseada em SNs de trabalhos que abordaram a identificação/extração de SNs. Os trabalhos selecionados eram formados por teses, dissertações, trabalhos de conclusão de curso e artigos.

Após o levantamento bibliográfico, foi definida a forma como seria feita a avaliação e comparação das ferramentas de extração automática de SNs. Para realizar a avaliação e comparação, foram extraídos manualmente os SNs presentes em 30 resumos de teses e dissertações, escolhidas aleatoriamente, de três áreas diferentes (Ciência da Computação, Direito e Nutrição). A escolha dos resumos de três áreas distintas foi proposital, pois queria-se observar como as ferramentas automáticas se comportariam.

Nessa pesquisa, o único programa capaz de extrair os SNs dos resumos foi o OGMA. O LX-Parser e o PALAVRAS apenas foram capazes de identificar, portanto foi necessária uma seleção manual de cada SN identificado por ambos. Após essa seleção, os SNs foram classificados de acordo com suas características e categorias, como: expressões que não constituem SNs; SNs compostos por palavras semelhantes às palavras-chave; SNs semelhantes às palavras-chave.

Silva (2014) utilizou as seguintes métricas para a computação dos resultados: a **taxa de acerto** (razão do total de expressões que constituem um sintagma nominal sobre o total de expressões extraídas por cada ferramenta); e a **taxa de revocação** (razão do número de expressões que são SNs extraídos pelos *softwares* sobre o número total de SNs extraídos manualmente).

Os resultados obtidos com a avaliação e comparação das ferramentas de extração e identificação de SNs indicaram que o LX-Parser, em alguns casos, mostrava uma performance superior aos outros. Entretanto, o número de identificação de falsos SNs foi muito mais alto no LX-Parser que no PALAVRAS e no OGMA. O PALAVRAS obteve uma taxa de erro de 6% que, segundo o autor, corrobora com as demais pesquisas que indicam o PALAVRAS como ferramenta com maior potencial para identificação de SNs.

O trabalho desenvolvido por Nascimento (2015) busca investigar a seleção de SNs que possuem valor descritivo no contexto da indexação automática por SNs, especificamente de resumos de teses e dissertações em língua portuguesa da área jurídica. Para desenvolver a pesquisa, o autor separou o trabalho em 5 etapas: 1) investigar a indexação automática por SNs; 2) verificar as características que compõem um sintagma nominal com valor de descritor documental; 3) buscar na literatura nacional metodologias e os critérios aplicados por cada metodologia para a seleção de SNs; 4) planejamento do experimento; 5) avaliação dos critérios de seleção.

O experimento realizado na pesquisa consiste na aplicação de critérios abordados na literatura científica sob os SNs extraídos dos 30 resumos de teses e dissertações da área jurídica, para assim mensurar e avaliar os critérios quanto a seleção de SNs descritores.

Para realizar essa avaliação, Nascimento (2015) comparou um conjunto de SNs, extraídos com o *software* PALAVRAS, com um conjunto de palavras-chave, identificando dessa forma os SNs descritores. Após a identificação dos SNs descritores é feita a aplicação dos critérios de seleção, dessa forma era identificado o comportamento de cada critério (Quadro 13) ao selecionar ou descartar um SN descritor. Para avaliar a utilidade e eficiência de cada critério o autor elaborou cálculos de revocação e precisão para cada um deles.

Quadro 13 – Categorias e critérios

Categorias	Função	Critérios
Eliminação	Eliminar SNs que não continham as palavras-chave (não descritores).	• Descarte de SNs com Numerais; • Descarte de SNs com pronomes como núcleo; • Descarte de SNs que iniciam com advérbios; • Descarte de SNs categorizados como Stopwords
Adição	Possibilita a identificação de SNs que contém as palavras-chave (descritores).	• Detecção de SNs contidos em SNs maiores por meio da remoção sucessiva de Adjetivos • Identificação de SNs por meio de Conjunção entre Adjetivos.
Ordenação	Ordenar os SNs por importância permite a identificação de SNs descritores dentro de determinados pontos de corte definidos através da análise do comportamento de cada critério de ordenação.	• Frequência de ocorrência dos SNs no Texto/Documento e Frequência Normalizada. • Frequência de ocorrência dos SNs no Conjunto de Documentos – Frequência Inversa de Documento – (IDF)
Nível	Permite verificar relação do nível do SN com a sua capacidade de funcionar como descritor documental.	• Estrutura e Nível do SN
Posição	Permite verificar se o posicionamento dos SNs no documento influência no resultado dele ser um descritor ou não.	• Posição do Sintagma Nominal no Documento

(**Fonte:** Nascimento (2015))

Os resultados obtidos pelos critérios foram diversos, mostrando que em alguns casos os resultados de revocação e precisão eram melhores que em outros, classificando cada um deles como útil o não. Os resultados dos critérios podem ser observados separadamente no quadro abaixo:

Quadro 14 – Resultado dos critérios

Critério	Resultado
Descarte de SNs com numerais	Resultados não tão satisfatórios, pois apesar de possuir uma revocação alta, a sua precisão era baixa.
Descarte de SNs com pronome como núcleo	Se mostrou útil, apesar de ter sido pouco aplicado devido ao número baixo de SNs que possuíam pronome como núcleo.
Descarte de SNs iniciados por advérbios	Obteve uma média de revocação alta e uma média de precisão baixa. Entretanto, esse critério não se mostrou útil, já que houve casos de eliminação de SNs considerados descritores.
Descarte de SNs categorizados como *stopwords*	Se mostrou bastante útil, obtendo uma média de revocação alta e de precisão acima da média alcançada sem a aplicação de nenhum critério.
Detecção de SNs contidos em SNs maiores por meio da remoção sucessiva de adjetivos,	Resultado não tão satisfatório, pois obteve médias de revocação e precisão baixa, portanto não foi útil para a seleção.
Identificação de SNs por meio de conjunção entre adjetivos	Esse critério não se mostrou útil para a seleção, pois obteve medias de revocação e precisão muito baixas.
Nível	Apesar de não ter apresentado uma taxa de revocação alta, a sua taxa de precisão foi acima da média. Esse critério se mostrou bastante pertinente, pois percebeu-se que grande parte dos SNs descritores se encontravam em SNs de nível 2 ou mais.
Posição/Ordem dos SNs	Esse critério foi classificado como útil, pois apesar de obter uma revocação baixa, teve uma precisão acima da média. Portanto, percebeu-se que essa relação se devia ao fato de que os primeiros SNs foram obtidos do título dos documentos.
Frequência de Ocorrência dos SNs no texto/documento e Frequência Normalizada	Concluiu-se que esse critério foi útil, pois apesar de ter uma revocação baixa, a sua precisão foi acima da média.
Frequência de ocorrência dos SNs no conjunto de documentos – Frequência inversa de documento (idf)	Obteve resultados satisfatórios de precisão e revocação, portanto foi classificado como útil na seleção de SNs descritores.

(**Fonte:** Nascimento (2015))

Analisando as pesquisas citadas, é evidente a variedade de critérios desenvolvidos para a seleção e extração de SNs, e de critérios desenvolvidos para processos de classificação automática de documentos, indexação automática ou manual, extração de conceitos, etc. Observou-se que, em alguns casos, a utilização

de determinados critérios obteve resultados mais significativos que os outros, o que auxilia no desenvolvimento da área de estudo da pesquisa. No quadro 15 são organizados, cronologicamente, os critérios utilizados em cada pesquisa citada neste trabalho, facilitando a visualização dos critérios que foram mais utilizados.

Quadro 15 – Critérios de identificação, seleção e extração de SNs das pesquisas

Pesquisa	**Critério**
Souza (2005)	• Frequência de ocorrência dos SNs no texto/documento; • Frequência de ocorrência dos SNs no conjunto de documentos do corpus (IDF); • Estruturas e níveis dos SNs; • Stoplist/Stopwords de SNs não descritores; • SN presente em tesauro da área.
Maia (2008)	• Frequência de ocorrência dos SNs no texto/documento; • Frequência de ocorrência dos SNs no conjunto de documentos do corpus (IDF); • Estruturas e níveis dos SNs.
Souza e Raghavan (2014)	• Frequência de ocorrência dos SNs no texto/documento; • Frequência de ocorrência dos SNs no conjunto de documentos do corpus (IDF); • Estruturas e níveis dos SNs; • Frequência de ocorrência normalizada do SN em um documento.
Silva (2014)	• Descarte de SNs iniciados com preposição ou conjunção; • Descarte de SNs que possuem verbos na estrutura; • Descarte de SNs que possuem numerais sem função de determinantes;
Nascimento (2015)	• Descarte de SNs com numerais; • Descarte de SNs com pronome como núcleo; • Descarte de SNs iniciados por advérbios; • Descarte de SNs categorizados como stopwords; • Detecção de SNs contidos em SNs maiores por meio da remoção sucessiva de adjetivos; • Identificação de SNs por meio de conjunção entre adjetivos; • Nível; • Posição/Ordem dos SNs; • Frequência de ocorrência dos SNs no texto/documento e Frequência Normalizada; • Frequência de ocorrência dos SNs no conjunto de documentos – Frequência inversa de documento (IDF).

(**Fonte:** desenvolvido pelo autor.)

Ao verificar o quadro 15 é evidente que alguns critérios foram mais utilizados que outros, como foi o caso dos critérios de frequência dos SNs (absoluta, inversa e

normalizada). A maioria dos critérios relacionados a frequência obtiveram resultados satisfatórios, o que demonstra que quanto maior a ocorrência do mesmo SN no documento, maior a chance desse SN ser considerado um descritor.

Com relação ao critério de nível, pesquisas como a de Souza (2005) e de Nascimento (2015) destacaram a importância e o cuidado que se deve tomar ao selecionar SNs descritores com base em seu nível estrutural. Souza (2005) salientou que os níveis dos SNs (complexos ou não) apresentam potencialidades diferentes, podendo, em alguns casos, não representar de forma clara e objetiva o documento. Nascimento (2015) destaca em sua pesquisa que os SNs de nível 2 ou maior apresentaram melhores resultados como descritores dos documentos.

As pesquisas de Corrêa *et al*. (2011) e de Nascimento (2015) destacaram também a importância do critério de posicionamento para a seleção de SNs. Corrêa *et al*. (2011) indagaram a possibilidade da utilização do critério de posicionamento dos SNs, da mesma forma que os sistemas de indexação utilizam para palavras isoladas. Esse critério foi aplicado e classificado como útil na pesquisa desenvolvida por Nascimento (2015), onde foi percebido que SNs que apareciam no início do texto, em alguns casos, eram classificados como SNs com potencial descritivo.

2.4.3 Normalização dos Sintagmas Nominais

O processo de normalização dos SNs é bastante delicado e complexo. Existem formas já propostas de extração e seleção de SNs que também se qualificam como critérios, princípios, diretrizes ou heurísticas para a normalização dos SNs.

Como a função principal do trabalho é apresentar uma proposta de normalização de SNs, nessa seção serão discutidas as principais pesquisas que abordaram essa temática e que contribuíram diretamente para a metodologia de normalização de SNs adotada nesse trabalho.

Uma pesquisa importante para o processo de normalização de SNs foi a de Lopes (2012), que tinha como objetivo propor, por meio de SNs, um processo de extração de termos que seriam conceitualmente relevantes.

Lopes (2012) classifica as suas heurísticas como sendo de extração, mas o papel que elas desempenham nos SNs também as qualifica como uma heurística de normalização de SNs.

No desenvolvimento de sua pesquisa, Lopes (2012) utilizou cinco *corpora* das seguintes áreas científicas: Geologia, Modelagem Estocástica, Mineração de Dados, Processamento Paralelo e Pediatria. Todavia, apenas o *corpus* de Pediatria foi utilizado na avaliação de todas as etapas propostas pela autora, pois era o único que tinha uma lista de termos de referência, os quais foram utilizados para comparação com os termos extraídos.

Lopes (2012) dividiu a sua pesquisa em 4 etapas, foram elas: 1) Extração de termos; 2) Ordenação de termos; 3) Identificação de conceitos; e 4) Aplicação dos termos e conceitos extraídos. A ferramenta utilizada para a extração dos termos foi o ExATO$_{LP}$ (Extrator Automático de Termos para Ontologias em Língua Portuguesa).

Após o processo de extração dos SNs, Lopes (2012) apresentou um conjunto de heurísticas, baseadas em análises linguísticas, com o objetivo de ajustar, reestruturar, incluir, excluir e refinar os SNs extraídos. As heurísticas desenvolvidas por Lopes (2012) foram:

Quadro 16 – Heurísticas de extração de sintagmas nominais

Heurísticas de ajuste	**1. Remoção de artigos que aparecem no início dos SNs**.
	2. Remoção de todos os artigos encontrados em um SN.
	3. Remoção dos pronomes que se encontram no início dos SNs – apenas aplicado quando o pronome não é o núcleo do sintagma nominal.
	4. Remoção de pronomes que se encontram em qualquer posição de um SN – propõe a remoção de todos os pronomes que não são núcleo do sintagma nominal. Exemplo: "o objetivo de seu movimento" passa a ser "o objetivo de movimento".
Heurísticas de descarte	**1. Descarte de SNs que contêm numerais**
	2. Descarte de SNs que contenham símbolos.
	3. Descarte de SNs que possuem como núcleo um "pronome" – quando ocorre anáfora.
	4. Descarte de SNs que iniciam com "advérbios" – Exemplo: "mais frequente".
Heurísticas de inclusão	**1. Detecção de SNs contidos em SNs maiores através da remoção sucessiva de adjetivos**, por exemplo, na frase: "Estudos realizados mostram o perigo de doenças virais hemorrágicas", em um processo básico de extração identificaria somente os seguintes SNs: "Estudos realizados", "perigo de doenças virais hemorrágicas" e "doenças virais hemorrágicas", no entanto, com a aplicação dessa heurística seriam gerados termos adicionais pela remoção dos adjetivos (ou verbos no particípio passado) ao fim de cada termo, gerando assim: "Estudos realizados", "Estudos", "perigo de doenças virais hemorrágicas", "perigo de doenças virais", e assim sucessivamente.
	2. Detecção de SNs replicados pelo uso de predicado múltiplo. Por exemplo, no SN : "pacientes idosos compram e tomam remédios mais caros". Assim os SNs "pacientes idosos" e "remédios mais caros" serão contabilizados duas vezes, pois apesar de aparecerem implícitos, eles ocorrem duas vezes.

	3. **Detecção de estruturas gramaticais múltiplas com o uso de conjunções, quando um substantivo é qualificado por mais de um adjetivo**. Por exemplo, na sentença: os pacientes idosos e obesos [...]. Têm-se dois SNs: “pacientes idosos” e “pacientes obesos”.

(**Fonte:** Lopes (2012, p.40))

Ao analisar as heurísticas propostas por Lopes (2012) é notório que as heurísticas de descarte e de ajuste devem funcionar em conjunto como critérios para a seleção de SNs que possivelmente podem representar conceitos.

Na heurística de descarte, a autora recomenda a remoção de sintagmas que possuam numerais, símbolos, dígitos, etc, pois os SNs que possuem essas características não devem ser classificados como descritores. Isso faz com que alguns SNs extraídos do título e resumo dos documentos sejam rapidamente excluídos e, após essa exclusão, são aplicadas as demais heurísticas nos SNs remanescentes.

A primeira e terceira heurística de inclusão proposta por Lopes (2012) estão relacionadas com um dos problemas relatados por Kuramoto (1995), chamado de “Sintagmas nominais escondidos em frases com fatoração”, que consiste em frases que possuem palavras procedidas por outra sequência de palavras ligadas por conjunções (“e”, “ou”, etc.). Por exemplo, na frase “As dívidas das empresas públicas e privadas” são identificados os seguintes SNs: “As dívidas das empresas públicas” e “As dívidas das empresas privadas”.

Com relação a relevância dos SNs, Lopes (2012) utilizou dois critérios: frequência do termo e divisão do termo no *corpus* de domínio em relação a *corpora*. Com base nesses critérios, foi estabelecido o índice tf-dcf (*term frequency*, *disjoint corpora frequency*).

A pesquisa de Lopes (2012) é bastante importante para o desenvolvimento deste trabalho, pois são apresentadas etapas de ordenação de termos e identificação de termos, cujos critérios serão citados e utilizados no decorrer deste trabalho.

Da pesquisa de Lopes (2012) podemos citar algumas heurísticas que são aplicadas para a normalização dos SNs e algumas heurísticas que são aplicadas para descarte dos SNs que podem ser aplicadas para a normalização.

As heurísticas de ajuste apresentadas por Lopes (2012) são classificadas como heurísticas de normalização, pois alteram a estrutura em que o sintagma nominal se apresenta, afim de melhorar a sua qualidade como descritor.

Todas as heurísticas de ajuste proposta por Lopes (2012) foram adaptadas e aplicadas para a proposta de normalização de SNs desenvolvida nessa pesquisa.

Além das heurísticas de ajuste, algumas das heurísticas de descarte também foram adaptadas para a normalização dos SNs. As heurísticas de descarte (1, 3 e 4) utilizadas por Lopes (2012) e que abordam numerais, advérbios e pronomes além de serem utilizadas como uma forma de descartar os SNs com menor valor descritivo, também podem ser aplicadas à normalização dos SNs, pois em alguns casos podem apresentar um valor descritivo para o documento.

Martins (2014) realizou uma pesquisa com o propósito de avaliar o uso de SNs como fontes de dados para um sistema automático de classificação de documentos textuais armazenados no formato digital. Para isso, ele dividiu a pesquisa em duas fases, a prospectiva e a consolidada. Em sua pesquisa, o autor demonstra a capacidade dos SNs serem utilizados como recursos na classificação temática de documentos de forma mais precisa e correta.

No decorrer da pesquisa, foram feitos diversos preparativos, como: 1) escolha do corpus; 2) preparação do corpus; 3) filtragem de conteúdo; 4) extração dos SNs; 5) correção dos SNs; 6) seleção dos SNs.

O *corpus* da pesquisa foi composto por 3 artigos sobre inteligência artificial. A fase de preparação do *corpus* consistia na remoção de possíveis imagens e tabelas presentes nos documentos, haja vista que a extração dos SNs é feita em documentos com o formato de texto simples (.TXT).

O processo de filtragem de conteúdo baseia-se na remoção de expressões que não contribuem como SNs, tais como: *Abstract*, Resumo, Palavras-chave, Introdução, Referências, etc.

Após a correção dos documentos, foi realizada a extração dos SNs através da ferramenta PALAVRAS e de um *script* escrito na linguagem de programação Perl (*Practical Extraction and Reporting Language*) pelo laboratório VISL. Após a extração, foi feita uma correção, que consistia em remover símbolos, espaços excessivos, aspas e qualquer outra representação que afetasse negativamente o processo de seleção dos SNs. Por conseguinte, percebeu-se que os SNs se repetiam, portanto foi feito também um processo de identificação dos sintagmas únicos.

Realizado todos os procedimentos para a preparação dos documentos e extração dos SNs, foi feita a aplicação do método preliminar, que utilizava as seguintes ferramentas: Oracle Virtual Box, Excel 2010, PALAVRAS, SVMLight, a linguagem PHP e o MS Visual Studio.

O método preliminar foi dividido em duas etapas. Na primeira etapa, considerando todos os sintagmas, é feita uma análise da semelhança estrutural que dois documentos podem ter, e o quanto essa semelhança influencia ao ponto de, talvez, serem classificados em uma mesma classe. Esta etapa é formada por quatro comparações: 1) Comparação direta entre sintagmas dos documentos; 2) Comparação após remoção dos quantificadores; 3) Comparação após o processo de stemming; 4) Comparação após a convergência de sinônimos. Na segunda etapa, o mesmo procedimento realizado na primeira etapa é feito novamente, porém com documentos de temas diferentes, a fim de avaliar se eles também poderiam ser classificados numa mesma classe, o que deveria ser evitado.

A comparação direta entre sintagmas dos documentos compara todos os sintagmas extraídos dos documentos, independentemente do número de vezes que tenham aparecido no documento.

Após a comparação direta, foi feita uma segunda comparação, dessa vez com os sintagmas sem os seus quantificadores (a, o, os, as, um, uma, uns, umas, alguns, algumas, numerais cardinais, numerais ordinais, pluralidade, pronomes universais, pronomes indefinidos, pronomes relativos e pronomes interrogativos), tendo o objetivo de avaliar o impacto causado pela mudança na estrutura de um sintagma. A remoção dos quantificadores dos sintagmas não gerou nenhuma melhora, logo não foi aplicada no método consolidado.

Em seguida, foi feita uma comparação dos SNs após o processo de *stemming*. O processo de *stemming* é responsável por reduzir palavras flexionadas ou derivadas ao seu tronco (*stem*). Martins (2014) exemplifica algumas palavras sem o *stemming*: “adaptação de os pesos de a rede” e após o *stemmming*: “adaptação pes red”.

Os resultados obtidos com a aplicação do processo de *stemming* foram satisfatórios, houve uma diminuição da ordem de 30% do tamanho do documento e melhora em relação ao uso de palavras no plural e no singular e, às vezes, no tempo verbal.

Percebe-se que as comparações feitas por Martins (2014), onde os SNs tiveram os seus quantificadores removidos; e os SNs após o processo de *stemming,* se qualificam como diretrizes de normalização de SNs, pois os SNs foram alterados com o objetivo de melhorarem os resultados obtidos.

Depois do processo de *stemming*, foi feita uma comparação após a convergência de sinônimos (processo que realiza a troca de algumas palavras por

seus sinônimos). A lista de sinônimos utilizada pelo autor foi retirada do *site* <sinonimos.com.br>. Com a comparação, notou-se que diversos fatores influenciam no processamento dos SNs, tais como: 1) não tratamento; 2) dicionário utilizado; 3) plural e singular; 4) não identificação de gêneros diferentes.

Ao término da primeira etapa, é realizada a segunda etapa, a qual consiste no desenvolvimento de uma contraprova, para validar os resultados. Nessa contraprova foram utilizados três documentos de um domínio diferente do utilizado anteriormente.

Os resultados obtidos na pesquisa de Martins (2014) indicaram que a utilização do processo de *stemming* foi vantajosa, se comparada com a utilização do próprio sintagma no momento de treinar a máquina de classificação automática. Martins (2014) também cita que trabalhos que usaram apenas sintagmas puros tiveram, em média, 80% de classificação satisfatória, ao passo que com a utilização do *stemming* esse percentual foi de 100%.

Martins (2014) não levou em consideração a frequência de ocorrência dos SNs para ponderá-los ou pontuá-los. Todavia, realizou um levantamento da frequência de ocorrência dos SNs extraídos, e utilizou apenas os SNs que apareceram no mínimo sete vezes no documento, com objetivo de afirmar a importância desses dados métricos para pesquisas desse gênero. A justificativa para a utilização dos SNs que aparecem sete vezes ou mais no documento foi de que havia uma queda significativa na frequência do sintagma seguinte.

> Uma métrica importante é a quantidade de vezes em que determinados sintagmas ocorrem no corpo de um documento. Essa informação pode ser relevante para sistemas de classificação, já que sintagmas que se repetem ao longo de um ou mais documentos de um corpus podem frequentemente se tornar um descritor importante nesse conjunto. (MARTINS, 2014, p. 73)

Analisando as pesquisas citadas, é notório os inúmeros critérios e diretrizes utilizados para a seleção de SNs e que em alguns casos eles se qualificam como diretrizes para a normalização de SNs. Para facilitar a compreensão desses critérios que podem ser qualificados como critérios de normalização, é possível observá-los cronologicamente no quadro 17.

Quadro 17 – Critérios de normalização de SNs das pesquisas

Pesquisa	Critério
Lopes (2012)	• Remoção de artigos que aparecem no início dos SNs. • Remoção de todos os artigos encontrados em um SN. • Remoção dos pronomes que se encontram no início dos SNs • Descarte dos SNs que contêm numerais; • Descarte dos SNs que possuem como núcleo um pronome; • Descarte dos SNs que iniciam com advérbios.
Martins (2014)	• Remoção de quantificadores; • *Stemming*;

(**Fonte:** desenvolvido pelo autor.)

2.4.4 Software de indexação automática por SNs - PyPLN

Nas pesquisas que estudaram a aplicação de SNs no processo de indexação automática de documentos foram utilizados diversos *softwares* com as funções de etiquetagem, identificação, seleção ou extração de SNs.

Alguns desses softwares já são conhecidos e foram citados anteriormente junto com as pesquisas que as utilizaram, tais como: OGMA, PALAVRAS e o Lx-Parser. Todavia, para o desenvolvimento dessa pesquisa foi escolhido o PyPLN.

O PyPLN é resultado de uma pesquisa desenvolvida pela Escola de Matemática Aplicada, da Fundação Getúlio Vargas, e trata-se de uma Plataforma de Processamento de Linguagem Natural desenvolvida em linguagem de programação Python e que pode ser acessada através do site http://pypln.org/.

A escolha do PyPLN se deve a dois fatores: 1) O PyPLN usa o parser PALAVRAS, que tem melhor desempenho na extração de SNs; 2) E através dele é possível realizar diversas análises linguísticas em textos em língua portuguesa, dessa forma possibilitando a identificação e extração dos SNs que estão presentes no corpus da pesquisa.

2.4.5 Avaliação da Indexação Automática por SNs

No processo de indexação, percebemos, na seção 2.2.2 (Avaliação da indexação), a existência de diversos critérios que possibilitam a avaliação da indexação manual, como: revocação, precisão e coerência. Alguns dos critérios

utilizados para avaliar a indexação manual também podem ser utilizados para avaliar a indexação automática.

Um critério importante para a avaliação da indexação automática é citado por Moens (2000), o critério de relevância. Através desse critério é possível se obter um *feedback* do usuário quanto ao processo de indexação automática realizado pelo sistema. O autor destaca que a avaliação da indexação automática possibilita a melhora do processo de recuperação de informação e do desempenho dos sistemas de recuperação de informação.

Hlava (2002) também apresenta alguns critérios que possibilitam avaliar sistemas de indexação automática, sendo um deles a análise dos índices selecionados automaticamente. Os índices analisados se classificam em três classes: 1) os que coincidem com os escolhidos pelo indexador humano; 2) os escolhidos pelo indexador e ignorado pelo processo automático; 3) e os que foram selecionados pelo computador, mas ignorados pelo indexador humano. Segundo a autora, através dessa análise, resultados que obtinham uma taxa de acerto superior a 85% eram considerados aceitos, sendo a taxa de acerto o percentual dos resultados que coincidem exatamente com o que o indexador humano aplicou.

Salton (1970), ao estudar o processo de indexação automática e manual, propôs em sua pesquisa uma fórmula capaz de avaliar os índices obtidos na indexação automática e manual e criou um coeficiente de avaliação comparativa entre as duas indexações. A fórmula utilizada para comparar a indexação automática e manual, foi:

$$q = \frac{c}{a + m - c}$$

Nessa fórmula, (q) equivale ao coeficiente de avaliação (comparativo entre as duas indexações), (c) é a quantidade de termos extraídos pelos dois sistemas que são comuns, (a) é a quantidade de termos selecionados pelo processamento automático e (m) é a quantidade de termos selecionados pelo processamento manual. Os valores do coeficiente variam entre 0 e 1, onde 0 significa que nenhum índice coincidiu e 1 significa que todos os índices, tanto automáticos como manuais, foram idênticos. Multiplicando-se o coeficiente por 100, resulta na porcentagem de aderência entre a indexação manual e a indexação automática.

A diferença da formula de consistência na indexação automática para a formula de avaliação da consistência da indexação manual são os critérios utilizados para a

avaliação. Na indexação manual são utilizados dois indexadores manuais (A e B), enquanto que na avaliação de consistência da indexação automática são utilizados um indexador manual e um indexador automático.

No âmbito da extração e seleção dos SNs, grande parte das pesquisas apresentam métodos diferentes de avaliação da qualidade da extração e seleção dos SNs, logo cabe ao pesquisador escolher ou desenvolver o melhor método para avaliar esses processos.

Geralmente, os métodos empregados pelos pesquisadores para avaliar os resultados, as técnicas e os critérios utilizados no processo de extração e seleção de SNs baseiam-se em: avaliações manuais dos resultados; análise dos resultados obtidos em diferentes sistemas; verificação de resultados obtidos manualmente e automaticamente; e comparação dos resultados obtidos por técnicas em cenários diferentes.

Como os métodos de avaliação da extração e seleção de SNs são determinados pelos pesquisadores em suas pesquisas, é relevante citar alguns trabalhos e os métodos de avaliação utilizados por eles.

Morellato (2007; 2010) e Miorelli (2001) utilizaram para comparar e mensurar níveis de satisfação os índices de equivalência dos termos entre os termos da extração automática e da extração manual, assim como Kuramoto (1995-2002).

Souza (2005) utilizou as palavras-chave como ferramenta de avaliação. O autor confrontava os SNs extraídos automaticamente, com as palavras-chave criadas pelos autores dos documentos, e após selecionados os descritores, eles eram comparados com as palavras-chave e os resumos dos documentos.

A forma utilizada por Santos (2005) para avaliar a extração dos SNs era baseada em equações de precisão e revocação na identificação dos SNs. Ao final da avaliação, o autor comparava os SNs identificados pela ferramenta PALAVRAS e os classificados pelo TBL (*Transformation-Based Learning* – Aprendizado Baseado em Transformações).

Na pesquisa de Maia (2008), o método de avaliação foi baseado em *softwares* (ED-CER e VISL) e tinha como objetivo avaliar a ferramenta desenvolvida pelo autor (OGMA). Esse método de avaliação permitia ao autor realizar correções e aprimorar a ferramenta da pesquisa.

Corrêa et al. (2011) realizaram a avaliação de forma manual, após submeter o corpus da pesquisa à ferramenta OGMA. A avaliação foi dividida em quatro aspectos:

1) verificar se os SNs extraídos realmente são SNs; 2) número de SNs relevantes; 3) número de SNs que contém as palavras-chave; 4) número de SNs que são iguais as palavras-chave. Durante esse processo de avaliação foram utilizados: o cálculo e análise dos percentuais de precisão na extração de SNs classificados como descritores; a taxa de erro com relação a extração de caracteres que não constituíam SNs; e o percentual de SNs extraídos que não são relevantes como descritores.

Lopes (2012, p.34) realizou a avaliação da extração e seleção dos SNs com base em uma lista, previamente estabelecida, de termos relevantes do domínio. Dessa forma, a autora era capaz de comparar (através de índices de precisão, revocação e medida F) a lista dos termos extraídos com a lista dos termos de referência. O índice de precisão é representado pela razão entre o número de termos extraídos obtidos da lista de referência e número de termos considerados. A revocação é a razão entre o número de termos da lista de extraídos e considerados presente na lista de referência e o tamanho da lista de referência. E a medida F expressa o equilíbrio entre os valores de precisão e revocação.

Silva (2014), avaliou a extração feita pelos *softwares* OGMA, PALAVRAS e LX-Parser, comparando as listas de SNs extraídos com cada *software*. As métricas utilizadas para comparar as listas de SNs extraídos foram: a taxa de acerto (razão do total de expressões que de fato constituem SN sobre o total de expressões extraídas por cada programa); e a taxa de revocação (razão do número de expressões que de fato são SNs extraídos pelos *softwares* sobre o número total de SNs extraídos pela técnica manual).

Nascimento (2015) utilizou valores numéricos para a avaliação dos critérios utilizados na seleção dos SNs que possuem potencial descritivo com relação ao documento. Nessa avaliação, foram analisados os valores de revocação e precisão obtidos com a aplicação de cada um dos critérios e comparados com os valores de revocação e precisão sem a aplicação de nenhum critério. Ao final, o autor indicou quais critérios se demonstraram úteis ou não para a seleção de SNs descritores em Maia e Souza (2010) e Martins (2014).

3 METODOLOGIA DA PESQUISA

A fundamentação teórica da metodologia utilizada nessa pesquisa foi baseada em autores como Chizzotti (1995), Gil (2010), Gonsalves (2007), Denzin e Lincoln (2006) e Marconi e Lakatos (2009), visto que desenvolveram trabalhos com o propósito de elucidar os procedimentos sistemáticos da metodologia científica, as técnicas de pesquisa e as metodologias de trabalho científico.

No que tange a natureza da pesquisa, ela se caracteriza como básica, posto que, através do estudo e pesquisa da literatura que abrange o campo de pesquisa e do estudo de caso desenvolvido, possibilita gerar novos conhecimentos sobre a indexação automática por SNs e, em especial, a respeito da normalização de SNs como descritores para a indexação automática.

No tocante ao objetivo da pesquisa, trata-se de uma pesquisa exploratória, pois a pesquisa exploratória possibilita uma maior familiaridade com o problema, visando torná-lo mais explícito ou construir hipóteses (GIL, 2010, p.41). E isso condiz com o fato da pesquisa aqui desenvolvida explorar a indexação automática por meio de SNs, com o objetivo de propor uma normalização para os SNs extraídos pela ferramenta PyPLN.

Com relação a abordagem do problema, a pesquisa é de caráter quantitativo, pois são apresentadas análises e métricas durante o estudo de normalização de SNs.

Segundo Gonsalves (2007, p.68), pesquisas de cunho quantitativo tentam esclarecer as causas, através da aplicação de métricas (estatística) e da criação de hipóteses. Por isso, no decorrer dessa pesquisa são trabalhados valores estatísticos acerca dos SNs extraídos e dos SNs normalizados em termos canônicos.

Quanto aos procedimentos, foi utilizada a pesquisa bibliográfica, pois através dela é possível se familiarizar com que foi produzido e discutido sobre a temática. A pesquisa bibliográfica foi realizada para alcançar os seguintes objetivos: compreensão do estado da arte; identificação dos fundamentos teóricos; análise dos critérios e metodologias utilizadas nas pesquisas; desenvolvimento de critérios de normalização dos SNs.

Através da pesquisa bibliográfica é possível identificar informações importantes acerca do objetivo proposto, contribuindo para a formação de um conhecimento mais aprofundado sobre a indexação, extração, seleção e normalização de SNs. Para o levantamento bibliográfico, foram considerados alguns contextos que envolvem o

tema, tais como: evolução dos conceitos de indexação; metodologias utilizadas para estudo de indexação automática baseadas em SNs; metodologia utilizadas para seleção, extração e normalização de SNs; ferramentas de extração de SNs.

Para o levantamento bibliográfico foram utilizados recursos físicos (bibliotecas) e digitais (*web*). Os principais tipos de documentos utilizados foram: dissertações, teses, livros e artigos científicos. Para a recuperação desses documentos foram utilizadas as seguintes bases de dados:

- Brapci;
- SciELO;
- PERI;
- Google Acadêmico;
- BDTD-IBICT;
- BDTD-UFPE;
- Portal de Periódicos da Capes.

Durante as pesquisas nessas bases dados, foram utilizados diversos termos, tanto em português como em inglês, e diversas combinações entre eles (estratégias de buscas) para o desenvolvimento da bibliografia. Alguns desses termos foram:

Quadro 18 – Lista de Termos

Sintagma nominal	Indexação automática	Indexação manual	Extração	Seleção
Avaliação	Indexação Manual	Sintagmas	Normalização	Critérios
PLN	Processamento de linguagem natural	PyPLN	Recuperação da Informação	Noun Phrases
Automatic Indexing	Extraction	Selection	Evaluation	Criteria

(**Fonte:** desenvolvido pelo autor.)

Cada trabalho recuperado nas bases de dados citadas, teve suas referências analisadas, pois algumas referências presentes nas pesquisas abordavam a mesma temática e contribuíam com informações pertinentes ao tema.

Esta pesquisa se classifica também como uma pesquisa empírica e utiliza como método o estudo de caso. Segundo Chizzotti (1995, p. 102), através da coleta e registro de dados é possível criticar ou avaliar analiticamente a experiência, visando tomar decisões ou propor ações. A pesquisa é classificada dessa forma, pois é realizado um experimento com o objetivo de avaliar uma metodologia para a normalização de SNs em termos canônicos para a sua utilização na indexação automática. Esse experimento procura validar critérios que possibilitem a normalização dos SNs extraídos através do *software* PyPLN, e para isso são utilizados os SNs relevantes extraídos dos títulos e resumos dos 60 documentos presentes no *corpus* de Souza (2005), os quais serão utilizados na pesquisa.

O método proposta na dissertação foi composta por critérios fundamentados na CI e levou em consideração os princípios e conceitos da linguagem documentária como principal base científica.

A pesquisa consistiu em 7 etapas gerais, que podem ser observadas a seguir:

1. Escolha dos documentos – Seleção dos 60 documentos que compõem o corpus de Souza (2005);
2. Coleta e organização dos documentos – os títulos e resumos são separados e preparados para a etapa seguinte;
3. Extração dos SNs – Processo automatizado feito pelo *software* PyPLN nos documentos;
4. Formatação dos SNs – Formatação dos SNs que irão participar do teste de revocação.
5. Seleção dos SNs relevantes – Aplicação de teste de revocação para selecionar SNs mais relevantes para a normalização;
6. Aplicação dos critérios de normalização – momento em que os SNs são submetidos à normalização;
7. Análise dos SNs normalizados –Análise de qualidade dos SNs normalizados.

A seguir, é detalhada cada uma das etapas.

3.1 Escolha dos documentos

Foi selecionado um total de 60 documentos, os mesmos documentos que fazem parte do *corpus* de Souza (2005). A decisão de selecionar esse *corpus* documental se

deve ao fato dele já ter sido utilizado em diversas pesquisas como Kuramoto (1995, 2002), Souza e Raghavan (2006, 2014), e isso expressa a sua qualidade como um material experimental para a pesquisa.

Dos 60 documentos que compõem o *corpus*, 29 deles são publicações da descontinuada revista **DataGramaZero**, e 31 deles são da revista **Ciência da Informação**. Apesar da extinção da revista DataGramaZero, alguns de seus documentos podem ser recuperados em outras bases de dados. Os documentos de ambas as revistas se encontravam disponíveis em Formato de Documento Portátil (PDF – *Portable Document Format*) ou em Linguagem de Marcação de Hipertexto (HTML – *HyperText Markup Language*).

Todos os documentos utilizados para a normalização abrangem os mesmos campos científicos, que são: biblioteconomia, documentação e ciência da informação.

3.2 Coleta e Organização dos Documentos

A etapa de coleta e organização consiste em extrair o título e resumo dos documentos que se encontravam em formato PDF e HTML, e transferi-los para arquivos separados com o formato de texto simples (TXT) e, em seguida, separar as palavras-chave para serem utilizadas durante a seleção de SNs relevantes.

A extração dos títulos e resumos para o formato TXT foi feita manualmente, pois a plataforma de extração utilizada (PyPLN) trabalha apenas com esse formato de arquivo. Cada documento possuía seu próprio arquivo TXT. Seria possível utilizar todos os resumos e títulos em apenas um arquivo TXT, mas o PyPLN não retornaria os SNs separados por documentos, o que não é interessante para a pesquisa.

Percebeu-se, enquanto coletava os resumos, títulos e palavras-chave dos documentos, a existência de palavras estrangeiras e o uso de símbolos, e eles não foram retirados, pois, em alguns casos, eles eram importantes para o documento e também para avaliar como o PyPLN lidava com essas informações.

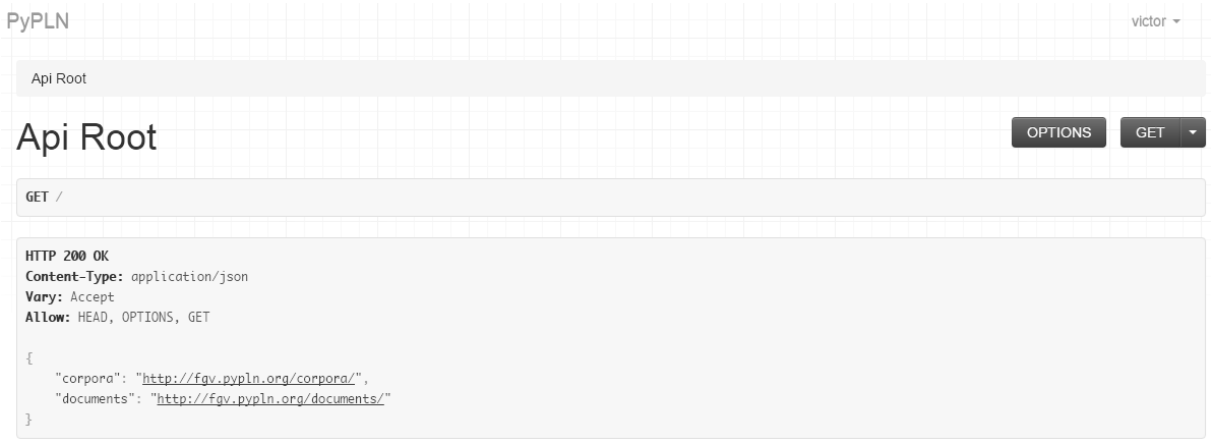
3.3 Extração dos SNs

Nessa etapa, todos os documentos, após separados em arquivos de texto simples, são submetidos à extração através da plataforma *web* do PyPLN (https://fgv.pypln.org/). Para utilizar o PyPLN é preciso fazer um registro, onde é solicitado um endereço de correio eletrônico (*e-mail*) e uma senha. Em seguida, um

e-mail contendo um *link* para a confirmação do registro é enviado e, ao clicar nesse *link*, o registro é efetuado. Após a realização do registro, é possível enviar para o PyPLN os arquivos, em texto simples, para que sejam extraídos os SNs.

A *interface* da plataforma se encontra em inglês e é bastante simples. Entretanto, inicialmente, devido a uma interface gráfica (Figura 5) não muito amigável, foi complicado compreender o seu funcionamento, e isso dificultou o envio dos documentos e baixar os resultados.

Figura 5 – Interface Gráfica PyPLN



(**Fonte:** Autoria nossa)

O processo de extração foi realizado em arquivos no formato TXT que continham os títulos e resumos de cada documento. Cada resumo continha, em média, 126 palavras, sendo o menor resumo com 32 palavras e o maior com 314 palavras, e a soma dos resumos tinha 7582 palavras. Cada título continha, em média 11 palavras, sendo o menor com 4 palavras e o maior com 21 palavras, e a soma de todos os títulos tinha 698 palavras.

Após o envio dos arquivos para o PyPLN, o processo de extração dos SNs levou, em média, 5 minutos por documento. Durante esse processo, 3 documentos precisaram ser reenviados, pois na página de resultados não constava nenhum SN. Contudo, após serem reenviados, conseguiu-se obter os SNs.

3.4 Formatação dos SNs

Uma filtragem de caracteres foi realizada, para eliminar os caracteres especiais retornados pela plataforma, como: “_” (sublinha), “*” (asterisco), aspas, chaves e sinais de pontuação (? ! , . : ;). O PyPLN retorna esses caracteres especiais como

forma de destacar o referente e o quantificador dos SNs. Por exemplo, no SN extraído: “o *valor de_ _o conhecimento”, o caractere sublinha é utilizado para denotar o quantificador do SN, e o asterisco para denotar o referente do SN.

Figura 6 – Caracteres especiais

```
Documento 01
"*Transferência da Informação :",
    "*análise para valoração de unidades de conhecimento .",
    "*valoração de unidades de conhecimento .",
    "*unidades de conhecimento .",
    "*conhecimento .",
    "o *valor de_ _o conhecimento",
    "_o *conhecimento",
    "_as mais *discutidas",
    "menos compreendidas *questões",
    "_os *estudos sobre a gestão de_ _o conhecimento .",
    "a *gestão de_ _o conhecimento .",
    "_o *conhecimento .",
```

(**Fonte:** Autoria nossa)

Percebeu-se, durante o processo de formatação, a presença de muitos SNs que podem ser considerados como sintagmas de significado vazio, pois não contribuem como descritores do conteúdo do documento. Entretanto, optou-se por levá-los à etapa seguinte, já que através da seleção de SNs relevantes esses sintagmas serão descartados.

Após a filtragem, 2786 SNs foram recuperados pelo PyPLN, e isso resulta numa média de, aproximadamente, 46 SNs por documento, sendo 100 o maior número de SNs extraídos por documento, referente ao documento 21, e 16 o menor número de SNs extraído, referente ao documento 58. No quadro a seguir são apresentados os números de SNs extraídos de cada documento:

Quadro 19 – Número de SNs por Documento

NÚMERO DE SNs POR DOCUMENTO							
Documento	**SNs**	**Documento**	**SNs**	**Documento**	**SNs**	**Documento**	**SNs**
DOC. 1	37	DOC. 16	42	DOC. 31	51	DOC. 46	61
DOC. 2	49	DOC. 17	41	DOC. 32	36	DOC. 47	20
DOC. 3	45	DOC. 18	28	DOC. 33	55	DOC. 48	43
DOC. 4	31	DOC. 19	36	DOC. 34	49	DOC. 49	46
DOC. 5	70	DOC. 20	47	DOC. 35	76	DOC. 50	45
DOC. 6	30	DOC. 21	100	DOC. 36	44	DOC. 51	27
DOC. 7	40	DOC. 22	95	DOC. 37	65	DOC. 52	52
DOC. 8	32	DOC. 23	42	DOC. 38	35	DOC. 53	70
DOC. 9	60	DOC. 24	33	DOC. 39	43	DOC. 54	76
DOC. 10	17	DOC. 25	65	DOC. 40	59	DOC. 55	50
DOC. 11	36	DOC. 26	43	DOC. 41	82	DOC. 56	49
DOC. 12	25	DOC. 27	49	DOC. 42	58	DOC. 57	29
DOC. 13	44	DOC. 28	48	DOC. 43	41	DOC. 58	16
DOC. 14	53	DOC. 29	36	DOC. 44	48	DOC. 59	37
DOC. 15	26	DOC. 30	37	DOC. 45	41	DOC. 60	45
						MÉDIA	**46,45**
						DESVIO	**17,09**

(**Fonte:** desenvolvido pelo autor.)

3.5 Seleção de SNs Relevantes

Nessa etapa foi realizada uma análise da revocação das palavras-chave, tendo como base os SNs extraídos pelo PyPLN e as palavras-chave atribuídas pelos autores.

A revocação das palavras-chave é feita através do cálculo do número de palavras-chave recuperadas em SNs dividido pelo número do total de palavras-chave do documento.

O teste de revocação permitiu selecionar os SNs que possuem as palavras-chave dos autores, pois são os potenciais SNs descritores que demandam normalização.

Portanto, todos os SNs que possuíam em sua estrutura as palavras-chave designadas pelos autores dos documentos foram selecionados para participar do processo de normalização.

Através dessa etapa foi possível observar quais SNs extraídos pelo PyPLN possuíam semelhanças com as palavras-chave dos autores, e isso atribuía a eles

mais valor uma vez que as palavras-chave são consideradas descritores do conteúdo do documento

No total, 452 SNs foram selecionados.

3.6 Proposição e aplicação do método de Normalização de SNS

Essa é a etapa mais importante do trabalho, pois ela trata do objetivo principal da pesquisa. Nela, os SNs selecionados são submetidos a diversos critérios que causarão alterações em sua estrutura a fim de transformá-los em termos canonizados.

Os critérios foram selecionados e desenvolvidos com base na pesquisa bibliográfica, onde foi possível analisar os critérios e as metodologias utilizadas pelos pesquisadores em seus trabalhos, e isso ajudou a desenvolver a proposta de normalização dos SNs.

A proposta de normalização de SNs consiste de duas etapas que levaram em consideração a preservação da validade ou da estrutura correta do SN, portanto nenhum SN deixou de ser um SN.

A principal regra limitadora da aplicação das etapas e dos critérios propostos para a normalização dos SNs seria a preservação do sentido expressado pelos SNs, pois esta é a característica que define o que são os sintagmas e não alterar a palavra-chave do autor.

A primeira etapa é composta por regras desenvolvidas com o objetivo de minimizar os SNs extensos, aproximando-os das palavras-chave dos autores e de termos descritivos. As regras (Quadro 20) foram desenvolvidas com base na análise de SNs, na identificação de classes gramaticais, e em expressões comuns.

Quadro 20 – Regras da Etapa 1

Regras	
Expressão Comum (Alguns + Pontos)	Expressão Comum (Por + Meio + Preposição)
Expressão Comum (Ante + Artigo)	Expressão Comum (Por + Parte + Preposição)
Expressão Comum (Após + Artigo/Preposição)	Expressão Comum (Possibilidade + Preposição)
Expressão Comum (Artigo + Análise + Preposição)	Expressão Comum (Preposição/Artigo + Planejamento + Preposição)
Expressão Comum (Artigo + Área + Preposição)	Expressão Comum (Pronome Demonstrativo + Contexto)

Expressão Comum (Artigo + Campo + Preposição)	Expressão Comum (Proposta + Preposição)
Expressão Comum (Artigo + Demanda + Atual + Preposição)	Expressão Comum (Relação (ões) + Entre)
Expressão Comum (Artigo + Foco + Preposição)	Expressão Comum (Relacionado (a) (s) + Artigo)
Expressão Comum (Artigo + Forma(s) + Preposição + Verbo)	Expressão Comum (Sobre + Artigo)
Expressão Comum (Artigo + Função + Preposição)	Que + Pronome Pessoal + Verbo
Expressão Comum (Artigo + Futuras + Pesquisas)	Que + Verbo
Expressão Comum (Artigo + Impacto + Preposição)	Que + Verbo + Advérbio;
Expressão Comum (Artigo + Importância + Preposição)	Que + Verbo + Preposição
Expressão Comum (Artigo + Linha + Preposição)	Que + Verbos
Expressão Comum (Artigo + Noção + Preposição)	Que + Verbos + Preposição
Expressão Comum (Artigo + Origem + De)	Preposição (Entre) + Artigo
Expressão Comum (Artigo + Papel + Preposição)	Preposição (Entre);
Expressão Comum (Artigo + Papel + Preposição)	Preposição (Pela (s))
Expressão Comum (Artigo + Pesquisa + Em)	Preposição (Sobre)
Expressão Comum (Artigo + Ponto + De + Vista + Preposição)	Preposição + Artigo
Expressão Comum (Artigo + Pressupostos + Teóricos)	Preposição + Preposição
Expressão Comum (Artigo + Problema + Preposição)	Preposição + Pronome Demonstrativo
Expressão Comum (Artigo + Própria (o) (s))	Preposição + Pronome Possessivo
Expressão Comum (Artigo + Reconhecimento + Preposição)	Preposição + Que
Expressão Comum (Artigo + Saída + Para)	Preposição + Que + Verbo
Expressão Comum (Artigo + Tarefa + de)	Preposição + Verbo
Expressão Comum (Artigo + Utilização + Preposição)	Conjunção + Artigo

Expressão Comum (Artigo/Preposição + Tema + De)	Conjunção + Preposição
Expressão Comum (As + Várias)	Conjunção + Pronome Possessivo
Expressão Comum (Aspectos + De)	Conjunção + Verbo (Particípio)
Expressão Comum (Através + De);	Verbo
Expressão Comum (Busca + De)	Verbo + Pelo (a)(s)
Expressão Comum (Cada + Tipo + De)	Verbo + Preposição
Expressão Comum (Capaz + De + Verbo)	Verbo + Que
Expressão Comum (Contida + Em)	Pronome Demonstrativo (Dessa (e))
Expressão Comum (Das + Principais)	Pronome Demonstrativo (Deste (a))
Expressão Comum (Definição + De);	Adjetivo Isolado
Expressão Comum (Diversos + Tipos + Preposição)	Sobre + Artigo
Expressão Comum (Enfoque + Especial)	Parênteses
Expressão Comum (Finalidade + Preposição)	Travessão
Expressão Comum (Gerado (a) (s) + Preposição);	
Expressão Comum (Implicação (ões) + Preposição)	
Expressão Comum (Iniciativa + De)	
Expressão Comum (Ligado + Artigo)	
Expressão Comum (Obstáculo + Artigo/Preposição)	

(**Fonte:** desenvolvido pelo autor.)

No total oitenta e nove (85) regras foram desenvolvidas e aplicadas na primeira etapa de normalização dos SNs. A ação tomada ao identificar as regras nos SNs era a de exclusão do elemento identificado pela regra e, se possível, a divisão do SN.

Por exemplo, no SN “os estudos sobre a gestão do conhecimento” foi identificada a regra “preposição + artigo” nos elementos “sobre a”. Sendo assim, os elementos que compõem a regra são excluídos e o SN é dividido em dois SNs. Portanto, do SN “os estudos sobre a gestão do conhecimento” surgiram dois SNs: “Os estudos” e “Gestão do conhecimento”.

A segunda etapa é composta por critérios que possuem o objetivo de eliminar elementos que não são importantes para os SNs. No total foram aplicados um total de sete critérios, um após o outro, durante o processo de normalização dos SNs. Os critérios utilizados foram:

A. Remoção de artigos no início e fim dos SNs;

B. Remoção de pronomes nos SNs;
C. Remoção de advérbios dos SNs;
D. Remoção de numerais;
E. Remoção de verbos;
F. Remoção de preposições e conjunções no início e no fim do SN;
G. Remoção do sufixo.

É importante que não seja alterado o significado dos SNs, uma vez que o principal motivo da escolha dos SNs como descritores é que eles, diferente das palavras isoladas, possuem significado.

Os critérios A, B e C foram baseados nos critérios de seleção de SNs propostos por Lopes (2012). Observou-se que esses critérios poderiam ser aplicados para a normalização dos SNs, uma vez que eles ajustam os SNs à uma forma mais normalizada.

A remoção dos artigos do início e fim dos SNs tem o intuito de extrair, de um SN, um termo candidato a conceito. Por exemplo, o SN “a produção da literatura”, após a remoção dos artigos, fica da seguinte forma: “produção de literatura”.

O critério de remoção de pronomes presentes nos SNs tem o mesmo objetivo do critério A, deixar o SN o mais próximo de um termo de indexação. Esse critério é aplicado apenas quando os pronomes não são o núcleo do SN. Dessa forma, o SN “seu direito à informação”, após aplicado o critério, ficaria da seguinte forma: “direito à informação”

Na pesquisa de Lopes (2012), o critério de remoção de advérbios foi utilizado para descartar SNs iniciados com advérbios. Todavia, na presente pesquisa muitos desses SNs foram descartados durante o teste de revocação. Ademais, alguns SNs que foram selecionados pelo teste de revocação possuíam advérbios, não no início, mas em outras posições, só que, se retirados, deixavam os SNs mais genéricos e próximos de um termo candidato a conceito. Como exemplo, o SN “uma área emergente dentro da ciência da informação”, após submetido a esse critério fica da seguinte forma: “uma área emergente da ciência da informação”.

O critério de remoção de numerais é justificado, pois os numerais (numerais, numerais cardinais, numerais ordinais) não contribuem para o significado do SN. Por exemplo, o SN “cinco programas de pesquisa”, após a remoção do numeral ficaria “programas de pesquisa” e essa forma fica mais próxima de um termo candidato a descritor.

O critério de remoção de verbo foi aplicado com o objetivo de avaliar a existência de verbos nos SNs que não fossem importantes para a descrição dos documentos. Todavia, esse critério não foi aplicado aos verbos que estivessem no particípio passado, pois os mesmos eram importantes para os SNs.

O critério de remoção de preposições e conjunções do início e fim dos SNs foi desenvolvido, pois foi identificada a presença desses elementos. Como esses elementos não são necessários para a descrição dos documentos, optou-se por remover aqueles que estivessem no início e no fim dos SNs e isso não alteraria o significado do SN.

Por fim, o critério de remoção de sufixo foi aplicado para comparação com os termos presentes no TBCI. Esse critério é o mais importante para a normalização dos SNs. Os outros critérios aplicados até o momento foram utilizados para simplificar e generalizar os SNs, dessa forma os SNs chegariam nesse último critério de uma forma mais simples e mais próximo de um termo de indexação.

O critério de remoção do sufixo consiste em, temporariamente, remover o sufixo das palavras dos SNs através do software Luke[8], para serem comparados manualmente com as palavras-chave e os termos do TBCI. Dessa forma, o SN que mais se assemelha de um termo presente no TBCI é considerado um SN normalizado, uma vez que o tesauro é considerado uma fonte de descritores canonizados, organizados e estruturados.

Para exemplificar esse critério, temos o seguinte SN que passou pelos outros critérios: “sistemas de recuperação”, com a aplicação do critério de remoção do sufixo, o SN fica, temporariamente, da seguinte forma: “sistem recuper”. Com isso, ao buscar manualmente esse SN no TBCI, a busca trará resultados mais abrangentes e, portanto, conseguirá recuperar descritores e suas variações de gênero (masculino e feminino) e número (singular e plural). Durante a aplicação desse critério, caso permaneçam no SN após a aplicação do método de normalização, são desconsideradas as conjunções e as preposições durante a busca do SN no TBCI, de forma que abranja ainda mais a busca de um termo descritor.

Com a aplicação da busca, os SNs que mais tiverem semelhança, que contenham ou que sejam idênticos a um termo presente no TBCI, serão considerados

[8] http://www.getopt.org/luke/

normalizados. Caso algum SN não se aproxime de algum termo presente no TBCI, esse SN será considerado um SN não normalizado.

A avaliação da normalização será feita com base no número de SNs que serão considerados normalizados, sem alterar o significado do SN e nem a palavra-chave. Para que os SNs sejam considerados normalizados não é necessário que ele tenha passado por algum dos critérios de normalização, pois alguns SNs não contemplam nenhum desses critérios, mas eles devem apresentar no critério de remoção de sufixo, algum termo descritor que esteja presente no TBCI.

4 ANÁLISE DOS RESULTADOS

Nesta seção são apresentados os resultados obtidos na aplicação das etapas do método proposto de normalização dos SNs.

A primeira análise a ser discutida no início desta seção é referente ao teste de revocação aplicado para a seleção dos SNs relevantes.

Após a análise do teste de revocação para a seleção dos SNs relevantes, apresenta-se a análise das etapas 1 e 2.

A avaliação de desempenho de cada critério é feita com base na não alteração do significado do SN e da preservação da palavra-chave do autor, ou seja, sem a perda do descritor. Aqueles que alterarem a palavra-chave identificada no SN ou que não apresentarem nenhuma alteração nos SNs selecionados para a normalização serão identificados através da avaliação.

4.1 Teste de Revocação

O teste de revocação teve o objetivo de selecionar os SNs mais qualificados para o processo de normalização e de analisar a utilização dos SNs extraídos do título e resumo como saída para sistemas de indexação automática. O teste da revocação utilizou as palavras-chave dos documentos presentes no *corpus* de Souza (2005). O cálculo de revocação foi feito com base no número de palavras-chave recuperadas a partir dos SNs extraídos pela plataforma PyPLN.

Os dados bibliográficos referentes aos documentos utilizados para a extração dos SNs podem ser observados no Apêndice A.

Esse teste de revocação selecionou os principais SNs retirados dos títulos e resumos dos documentos presentes no corpus de Souza (2005) e classificou-os como SNs de alto nível descritivo, já que possuíam em sua estrutura as palavras-chave dos documentos.

Após o teste de revocação, 4 documentos (documentos 15, 17, 49, 59) foram descartados, uma vez que nenhum de seus SNs foram selecionados para o processo de normalização, devido ao fato de não possuírem, no título e no resumo, as palavras-chave dos autores.

Ademais, o teste de revocação possibilitou analisar se os SNs extraídos apenas do título e resumo dos documentos servem como saída para os sistemas de indexação

automática. Através dos valores de revocação obtidos percebeu-se que é plausível a utilização desses SNs, porém com algumas ressalvas.

Após a extração dos SNs com o *software* PyPLN, e o cálculo da revocação das palavras-chave, foram obtidos os valores que podem ser observados no Quadro 21.

Quadro 21 – Resultado da Revocação

VALOR DA REVOCAÇÃO DAS PALAVRAS-CHAVES POR DOCUMENTO									
Documento	**Revocação**		**Documento**	**Revocação**		**Documento**	**Revocação**		**Documento**	**Revocação**
DOC. 1	33.3%		DOC. 16	20.0%		DOC. 31	60.0%		DOC. 46	83.3%
DOC. 2	100.0%		DOC. 17	0.0%		DOC. 32	66.7%		DOC. 47	33.3%
DOC. 3	40.0%		DOC. 18	25.0%		DOC. 33	80.0%		DOC. 48	66.7%
DOC. 4	60.0%		DOC. 19	40.0%		DOC. 34	66.7%		DOC. 49	0.0%
DOC. 5	60.0%		DOC. 20	25.0%		DOC. 35	60.0%		DOC. 50	50.0%
DOC. 6	42.9%		DOC. 21	20.0%		DOC. 36	50.0%		DOC. 51	80.0%
DOC. 7	20.0%		DOC. 22	66.7%		DOC. 37	50.0%		DOC. 52	80.0%
DOC. 8	100.0%		DOC. 23	66.7%		DOC. 38	100.0%		DOC. 53	60.0%
DOC. 9	80.0%		DOC. 24	66.7%		DOC. 39	60.0%		DOC. 54	14.3%
DOC. 10	40.0%		DOC. 25	50.0%		DOC. 40	100.0%		DOC. 55	50.0%
DOC. 11	25.0%		DOC. 26	75.0%		DOC. 41	100.0%		DOC. 56	20.0%
DOC. 12	50.0%		DOC. 27	33.3%		DOC. 42	80.0%		DOC. 57	57.1%
DOC. 13	33.3%		DOC. 28	42.9%		DOC. 43	100.0%		DOC. 58	71.4%
DOC. 14	33.3%		DOC. 29	100.0%		DOC. 44	66.7%		DOC. 59	0.0%
DOC. 15	0.0%		DOC. 30	66.7%		DOC. 45	80.0%		DOC. 60	100.0%
									MÉDIA	**55,0%**
									DESVIO	**28,3%**

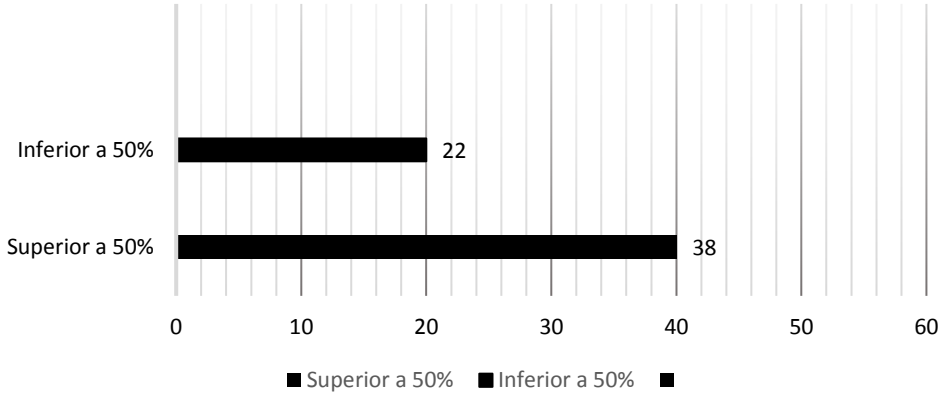
(**Fonte:** desenvolvido pelo autor.)

A média da revocação das palavras-chave, obtida através do teste realizado no *corpus*, foi de 55%. Portanto, mais da metade das palavras-chave indicadas pelos autores foram recuperadas através dos SNs extraídos automaticamente pelo PyPLN. Esse resultado é bastante positivo, pois indica que grande parte dos documentos possuem SNs no título e resumo que possibilitam a sua recuperação em um sistema de recuperação de informação. O valor apresentado poderia ser melhor se alguns problemas, como a utilização de caracteres especiais e utilização de palavras estrangeiras fossem tratados. Os problemas encontrados são descritos mais adiante.

Outro resultado interessante foi o desvio padrão de 28%. O desvio padrão significa a medida de variação, ou dispersão, obtida pela média da revocação, ou seja, ele mede a variabilidade dos valores em torno da média, quanto mais baixo o desvio padrão, os dados tendem a estar mais próximos da média. Percebe-se que o desvio padrão foi aproximadamente metade do valor da média, o que indica que as variações dos valores de revocação de palavras-chave foram próximos da média.

Dos 60 documentos, 22 deles (36,6%) tiveram um valor de revocação inferior a 50% (Gráfico 2) e, dentre estes, 4 (6,6%) não tiveram valor de revocação, pois não foram recuperadas palavras-chave nos SNs extraídos pelo PyPLN. Ademais, 38 documentos (63,3% dos documentos) apresentaram uma revocação igual ou superior a 50% das palavras-chave.

Gráfico 2 – Quantidade de documentos por nível de revocação



(**Fonte:** desenvolvido pelo autor.)

A baixa revocação das palavras-chave em alguns documentos foi causada por dois fatores de indexação.

O primeiro fator foi a ausência das palavras-chave no título e no resumo do documento. Essa ausência impossibilitava o *software* PyPLN de recuperar SNs que possuíssem as palavras-chave. No total, 4 documentos foram afetados por esse problema.

O segundo fator foi quanto ao formato da palavra-chave. Em alguns casos, o formato da palavra-chave se aproximava de algum SN extraído pelo PyPLN, mas como não era exatamente igual, não sendo considerado no cálculo de revocação. Como exemplo, podemos citar o caso do documento 15, nele a palavra-chave "*Ensino e pesquisa*" se aproximava do SN "*Relação Ensino-Pesquisa*".

Houve dois documentos que apresentaram as palavras-chave de forma diferenciada. Os documentos 16 e 25 apresentaram o uso do caractere especial (travessão) para especificar a área temática da palavra-chave. Por exemplo: "*Fomento à pesquisa – Ciência da Informação; CNPq – fomento à pesquisa em Ciência da Informação*", devido à grande extensão e à utilização do caractere especial (travessão) para separar o texto que compõe a palavra-chave, não havia a possibilidade do PyPLN extrair um SN equivalente à palavra-chave. Portanto foi

selecionado como palavra-chave para o teste de revocação, a palavra-chave que antecede o travessão, pois na indexação ela é a palavra-chave utilizada para a área temática especificada após o travessão.

Apesar de 22 documentos (36,6%) terem valor de revocação inferior a 50%, isso não é ruim, pois não significa que o documento não é recuperável através de um SRI. Todavia, significa que ele possui poucos SNs relevantes no título e resumo capazes de propiciar a recuperação do documento.

Se por um lado 36,6% dos documentos tiveram o valor de revocação abaixo de 50%, por outro, o número de documentos que tiveram uma revocação igual ou superior a 50% foi bastante satisfatório. No total, foram 38 documentos, ou seja, 63,3% dos documentos, sendo que 8 deles (13%) tiveram uma revocação de 100% das palavras-chave dos autores.

Ao analisar os resultados do teste de revocação dos 32 documentos com revocação superior a 50%, ficou evidente que a positividade desses resultados não se deve a quantidade de SNs extraídos, mas sim ao fato de que as palavras-chave apresentadas pelos documentos estavam presentes no resumo ou no título e foram extraídas como SNs pelo PyPLN.

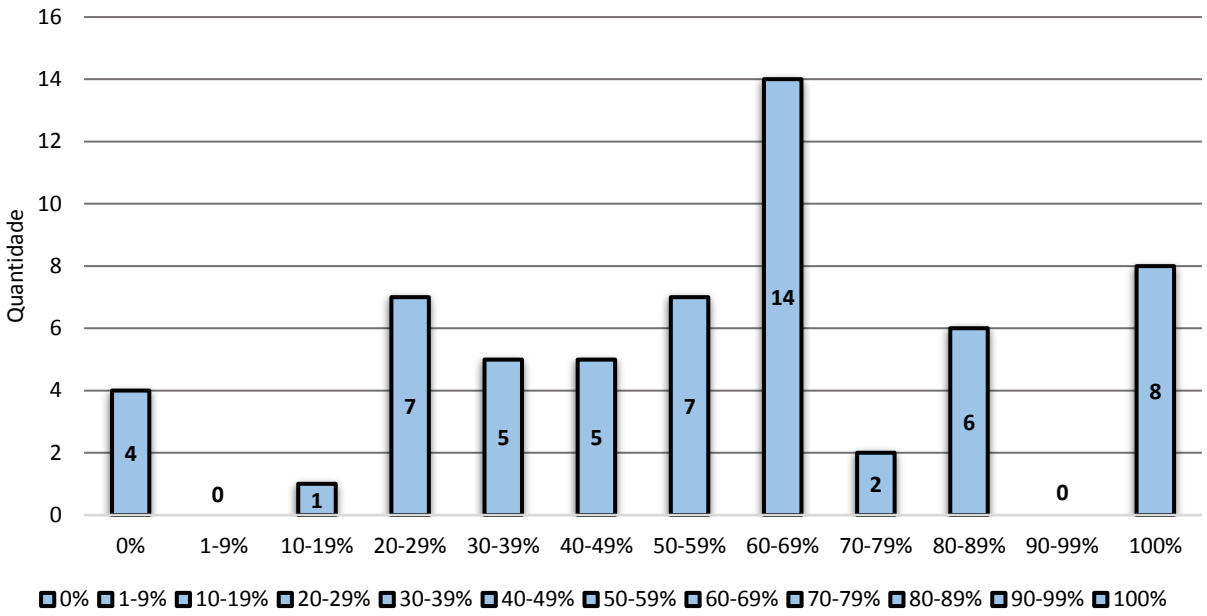
Nos 8 documentos que tiveram o valor de revocação de 100%, as palavras-chave utilizadas pelos autores não apresentavam o uso de nenhum caractere especial (travessão, aspas, etc.).

Em alguns casos, a presença de caracteres especiais no resumo ou no título do documento dificultou a extração do SN de forma correta pelo PyPLN. Por exemplo, no documento 42 existem duas palavras-chave: Literatura Cinzenta e Literatura Branca. Porém as palavras Cinzenta e Branca se encontravam entre aspas no texto e por isso o *software* não conseguiu extrair esses dois SNs, causando uma queda na revocação para esse documento.

Outro fator que impossibilitou a extração de alguns SNs foi o uso de termos estrangeiros. Como por exemplo, o documento 54 possui o SN “*Informacion Literacy”* e o PyPLN extraiu incorretamente cada palavra como um SN.

No geral, a revocação obtida foi bastante satisfatória, pois muitos documentos possuíram a possibilidade de serem recuperados, mesmo com alguns apresentando um baixo valor de revocação. Ao todo, 93% dos documentos (56) poderiam ser recuperados através de uma base de dados que utilizasse os SNs extraídos do título e resumo (Gráfico 3).

Gráfico 3 – Quantidade de documentos por faixas de valores de revocação



(**Fonte:** desenvolvido pelo autor.)

Enquanto 4 documentos não obtiveram nenhuma palavra-chave recuperada do total de 13 palavras-chave, 18 documentos com nível de revocação entre 10% e 49% tiveram 26 palavras-chave recuperadas do total de 86 palavras-chave; 30 documentos com nível de revocação entre 50% e 89% tiveram 93 palavras-chave recuperadas do total de 140 palavras-chave; e 8 documentos com nível de revocação de 100% tiveram um total de 27 palavras-chave recuperadas.

Após a análise do resultado da revocação, é feita uma avaliação dos SNs que foram selecionados através do teste de revocação com o intuito de afirmar se eles podem ser classificados como SNs de alto nível descritivo e, portanto, podem seguir para a etapa de normalização. Nessa avaliação são feitas algumas observações quanto aos SNs selecionados e descartados no teste de revocação, indicando quais foram os principais problemas e suas possíveis soluções.

Antes do teste de revocação, 2786 SNs foram extraídos dos documentos, resultando em uma média de, aproximadamente, 46 SNs por documento. Após o teste de revocação, o total de SNs remanescente foi de 452, ficando então com uma média de, aproximadamente, 8 SNs por documento.

Quadro 22 – Número de SNs Após Teste de Revocação

NÚMERO DE SNs APÓS TESTE DE REVOCAÇÃO							
Documento	**SNs**	**Documento**	**SNs**	**Documento**	**SNs**	**Documento**	**SNs**
DOC. 1	2	DOC. 16	1	DOC. 31	5	DOC. 46	21
DOC. 2	5	DOC. 17	0	DOC. 32	11	DOC. 47	4
DOC. 3	11	DOC. 18	3	DOC. 33	7	DOC. 48	9
DOC. 4	9	DOC. 19	6	DOC. 34	7	DOC. 49	0
DOC. 5	11	DOC. 20	1	DOC. 35	28	DOC. 50	3
DOC. 6	5	DOC. 21	1	DOC. 36	4	DOC. 51	15
DOC. 7	9	DOC. 22	15	DOC. 37	4	DOC. 52	7
DOC. 8	10	DOC. 23	11	DOC. 38	8	DOC. 53	7
DOC. 9	12	DOC. 24	17	DOC. 39	4	DOC. 54	6
DOC. 10	3	DOC. 25	5	DOC. 40	7	DOC. 55	8
DOC. 11	3	DOC. 26	3	DOC. 41	17	DOC. 56	2
DOC. 12	6	DOC. 27	4	DOC. 42	10	DOC. 57	14
DOC. 13	1	DOC. 28	9	DOC. 43	10	DOC. 58	11
DOC. 14	5	DOC. 29	11	DOC. 44	8	DOC. 59	0
DOC. 15	0	DOC. 30	15	DOC. 45	7	DOC. 60	14
						MÉDIA	**7,53**
						DESVIO	**5,54**

(**Fonte:** desenvolvido pelo autor.)

Ao analisar isoladamente cada documento, percebe-se que, em alguns casos, a redução no número de SNs foi bastante alta. Isso foi consequência dos mesmos problemas citados durante a análise dos valores de revocação: ausência da palavra-chave no título e resumo, presença de caracteres especiais no título ou no resumo ou na palavra-chave e a extensão de algumas palavras-chave.

Os problemas relativos a plataforma de extração (PyPLN) também contribuíram para o descarte indevido de alguns SNs. Esses problemas foram causados devido à presença de caracteres especiais e de termos estrangeiros no título ou resumo.

Para solucionar os problemas relacionados ao uso de caracteres especiais e termos estrangeiros é recomendado que o tratamento desses problemas seja inserido na plataforma de extração (PyPLN).

Quanto à ausência de palavras-chave no título e resumo dos documentos, é possível imaginar duas soluções. A primeira seria a reformulação, por parte dos autores, do título e resumo de seus trabalhos. E a segunda seria a utilização do texto completo, e não apenas do título e resumo dos documentos durante processo de extração dos SNs.

Por suposto, no geral, os SNs selecionados através do teste de revocação podem ser considerados os mais adequados para a aplicação da proposta de

normalização, pois apresentam em suas estruturas as palavras-chave que foram escolhidas pelos autores dos documentos.

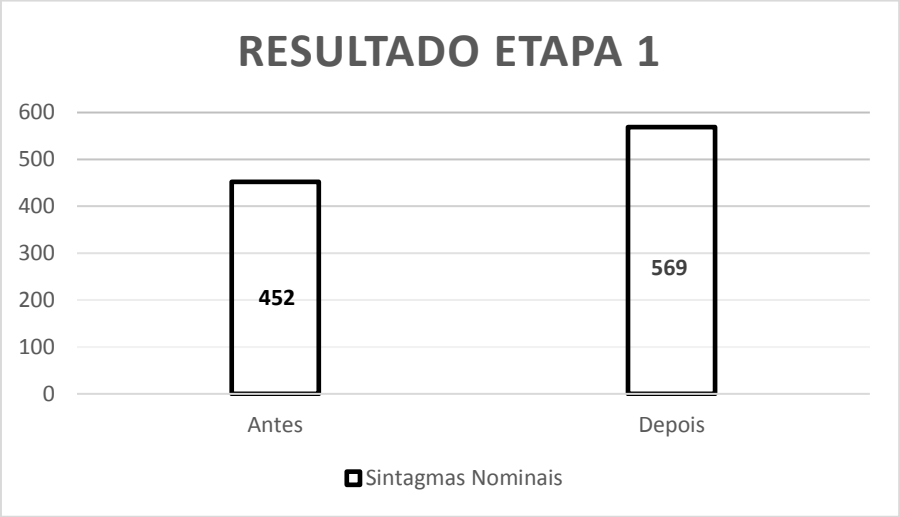
4.2 Análise da etapa 1

A etapa 1 teve o objetivo de minimizar os SNs extensos e aproximá-los de termos descritores. As regras desenvolvidas para a aplicação dessa etapa levaram em consideração a preservação da estrutura correta do SN, portanto nenhum SN deixou de ser um SN.

As regras da etapa 1 da normalização dos SNs foram criadas com base nas pesquisas de Lopes (2012) e Martins (2014).

Ao observar o gráfico 4, é possível notar que no início da etapa 1 existiam 452 SNs. Após a sua aplicação, o número de SNs passou a ser de 569, ou seja, um aumento de 117 SNs, aproximadamente 25,8%.

Gráfico 4 – Resultado da Etapa 1



(**Fonte:** desenvolvido pelo autor.)

As regras da etapa 1 foram aplicadas em 138 SNs e ocasionaram o surgimento de 254 novos SNs. Alguns desses SNs eram semelhantes a outros SNs já selecionados para a aplicação da proposta de normalização, mas isso não ocasionou a remoção desses SNs.

Para exemplificar a aplicação de um critério é indicado observar um breve exemplo. Antes da aplicação da etapa 1 temos o SN “um processo otimizado por agentes inteligentes”, ao analisar identificamos a regra de “Verbo + Preposição (por)”.

Ao aplicar a regra identificada da etapa 1 no SN, teremos o surgimento de dois novos SNs: “Um Processo” e “Agentes Inteligentes”.

No Quadro 23 é possível observar o número de SNs por documento antes (SNa) e depois (SNd) da aplicação da etapa 1.

No Apêndice B são apresentadas detalhadamente a aplicação de cada uma das regras nos SNs que foram divididos. É possível identificar através da descrição da regra a forma como o SN foi alterado após a sua aplicação.

Quadro 23 – Número de SNs na Etapa 1

NÚMERO DE SNs ANTES(SNa) E DEPOIS(SNd)												
Documento	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**
DOC. 1	2	3	DOC. 16	1	1	DOC. 31	5	6	DOC. 46	21	26
DOC. 2	5	9	DOC. 17	0	0	DOC. 32	11	15	DOC. 47	4	5
DOC. 3	11	11	DOC. 18	3	3	DOC. 33	7	8	DOC. 48	9	13
DOC. 4	9	13	DOC. 19	6	9	DOC. 34	7	7	DOC. 49	0	0
DOC. 5	11	13	DOC. 20	1	2	DOC. 35	28	33	DOC. 50	3	4
DOC. 6	5	5	DOC. 21	1	1	DOC. 36	4	4	DOC. 51	15	18
DOC. 7	9	10	DOC. 22	15	18	DOC. 37	4	4	DOC. 52	7	7
DOC. 8	10	14	DOC. 23	11	13	DOC. 38	8	20	DOC. 53	7	8
DOC. 9	12	12	DOC. 24	17	24	DOC. 39	4	4	DOC. 54	6	15
DOC. 10	3	3	DOC. 25	5	7	DOC. 40	7	8	DOC. 55	8	10
DOC. 11	3	3	DOC. 26	3	4	DOC. 41	17	18	DOC. 56	2	3
DOC. 12	6	6	DOC. 27	4	12	DOC. 42	10	14	DOC. 57	14	14
DOC. 13	1	1	DOC. 28	9	7	DOC. 43	10	20	DOC. 58	11	14
DOC. 14	5	5	DOC. 29	11	11	DOC. 44	8	10	DOC. 59	0	0
DOC. 15	0	0	DOC. 30	15	15	DOC. 45	7	7	DOC. 60	14	17

(**Fonte:** desenvolvido pelo autor.)

A aplicação da etapa 1 se mostrou bastante positiva, pois, como esperado, ela conseguiu minimizar diversos SNs que, ainda eram muito extensos para serem apropriadamente normalizados.

De acordo com o Quadro 24, houve uma modificação de 8 SNs de nível 1; 37 SNs de nível 2; 41 SNs de nível 3; 24 SNs de nível 4; 11 SNs de nível 5; 9 SNs de nível 6; 3 SNs de nível 7; 2 SNs de nível 8; 1 SN de nível 9; 1 SN de nível 10 e 2 SNs de nível 11.

Quadro 24 – Número de SNs Alterados por Nível na Etapa 1

Nível	1	2	3	4	5	6	7	8	9	10	11
Total	8	37	41	23	11	9	3	2	1	1	2

(**Fonte:** desenvolvido pelo autor.)

No quadro 25 é possível observar que os novos 254 SNs apresentaram os seguintes números: 125 SNs de nível 1; 68 SNs de nível 2; 32 SNs de nível 3; 16 SNs de nível 4; 6 SNs de nível 5; 7 SNs de nível 6 e 1 SN de nível 7.

Quadro 25 – Nível dos SNs Alterados na Etapa 1

Nível	1	2	3	4	5	6	7	8	9	10	11
SNs Antes	8	37	41	23	10	9	3	2	1	1	2
SNs Depois	125	68	32	16	6	7	1	0	0	0	0

(**Fonte:** desenvolvido pelo autor.)

Dos 254 novos SNS, 101 não possuem nenhuma das palavras-chave dos autores, porém continuarão adiante no processo de normalização dos SNs, dessa forma será possível verificar se as aplicações dos critérios geram SNs válidos.

Ao analisar os SNs do Apêndice B resultantes da etapa 1, bem como o Quadro 25, é possível perceber uma diminuição no número de SNs extensos, e com isso diversos SNs conseguiram se aproximar ainda mais das palavras-chave dos autores do documento. Portanto, é possível afirmar que o objetivo da etapa 1 foi atingido.

Todas as regras que foram definidas nessa etapa foram aplicadas manualmente sem dificuldades, sendo assim é possível que essas regras sejam utilizadas para a aplicação automática por um sistema.

Após a análise dos resultados apresentados pela etapa 1, é iniciada a análise dos resultados da etapa 2. Para facilitar a compreensão, os resultados da etapa 2 foram analisados separadamente, seguindo a ordem de aplicação dos critérios.

4.3 Remoção de artigos do início e fim dos SNs

A aplicação do critério de remoção de artigos do início e fim dos SNs extraídos do *corpus* tem o objetivo de verificar se ele é necessário para o processo de normalização dos SNs. Este critério foi utilizado de forma similar pelo trabalho de Lopes (2012) para a seleção de conceitos.

Em sua pesquisa, Lopes (2012) fez a aplicação do mesmo critério em duas partes: primeiro, na remoção de artigos no início do SN; e, em seguida, a remoção de todos os artigos do SN. Contudo, na presente pesquisa foi feita a remoção dos artigos presentes apenas no início e no fim do SN.

Lopes (2012) identificou, através da heurística de remoção de artigos no início do SN, o ajuste de cerca de 43% dos SNs, enquanto que através da heurística de ajuste de remoção de todos os artigos do SN houve um ajuste de cerca de 49%.

Como esperado, assim como na pesquisa de Lopes (2012), a aplicação desse critério foi a mais impactante, pois de 569 SNs aprovados para a etapa 2, 231 SNs, cerca de 40,5%, tiveram alterações devido à existência de artigos irrelevantes em sua estrutura.

Para melhor compreender a aplicação desse critério é indicado a observação de alguns exemplos. Observe os seguintes SNs: “A ciência da informação” e “Às necessidades”, em ambos SNs temos a presença de artigos definidos, sendo um deles craseado. Por se qualificarem no critério de remoção de artigos no início e no fim do SN, eles terão os artigos removidos e ficarão da seguinte forma: “Ciência da informação” e “Necessidades”.

No Apêndice C estão disponibilizados todos os SNs que sofreram alterações devido a aplicação da regra de remoção de artigos. Nele é possível perceber a estrutura dos SNs antes e depois da aplicação e assim evidenciar que todos os SNs foram devidamente alterados.

Durante a aplicação desse critério nenhuma das palavras-chave dos autores que estavam presentes nos SNs foi alterado, portanto foi mantido o princípio de preservar a palavra-chave dentro do SN e do significado do SN.

Através do Quadro 26 é possível visualizar, por documento, o número de SNs a normalizar e o número de SNs que sofreram alterações devido à aplicação do critério de remoção de artigos do início e fim dos SNs.

Os artigos definidos foram os mais excluídos em comparação com os indefinidos, e todos os artigos removidos estavam posicionados no início do SNs. No total, 231 artigos foram removidos dos SNs: 211 artigos definidos e 20 artigos indefinidos.

O principal motivo para a exclusão desses artigos foi que, além de não influenciarem para o sentido da frase, eles não são considerados em diversos sistemas de busca e pela regra de desenvolvimento de tesauros. Dessa forma, não há necessidade da permanência desses artigos.

Quadro 26 – Número de SNs no 1º Critério da Etapa 2

NÚMERO DE SNs ANTES(SNa) E ALTERADOS (SNd)												
Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**
DOC. 1	3	3	DOC. 16	1	1	DOC. 31	6	3	DOC. 46	26	12
DOC. 2	9	5	DOC. 17	0	0	DOC. 32	15	8	DOC. 47	5	1
DOC. 3	11	8	DOC. 18	3	2	DOC. 33	8	1	DOC. 48	13	4
DOC. 4	13	10	DOC. 19	9	2	DOC. 34	7	1	DOC. 49	0	0
DOC. 5	13	4	DOC. 20	2	1	DOC. 35	33	7	DOC. 50	4	3
DOC. 6	5	5	DOC. 21	1	0	DOC. 36	4	2	DOC. 51	18	8
DOC. 7	10	5	DOC. 22	18	6	DOC. 37	4	0	DOC. 52	7	2
DOC. 8	14	6	DOC. 23	13	4	DOC. 38	20	15	DOC. 53	8	2
DOC. 9	12	5	DOC. 24	24	12	DOC. 39	4	0	DOC. 54	15	7
DOC. 10	3	2	DOC. 25	7	3	DOC. 40	8	4	DOC. 55	10	4
DOC. 11	3	1	DOC. 26	4	0	DOC. 41	18	5	DOC. 56	3	2
DOC. 12	6	2	DOC. 27	12	4	DOC. 42	14	7	DOC. 57	14	3
DOC. 13	1	1	DOC. 28	7	3	DOC. 43	20	8	DOC. 58	14	4
DOC. 14	5	1	DOC. 29	11	3	DOC. 44	10	5	DOC. 59	0	0
DOC. 15	0	0	DOC. 30	15	5	DOC. 45	7	5	DOC. 60	17	4

(**Fonte:** desenvolvido pelo autor.)

Outro dado interessante foi em relação ao número de SNs que sofreram alterações de acordo com seu nível. É possível observar no quadro 27 que os SNs de nível 1 e 2 foram os mais alterados, e isso condiz com os números, pois os SNs de nível 1 e 2 compõem a maioria dos SNs recuperados pelo PyPLN.

Quadro 27 – Nível dos SNs Alterados no 1º Critério

	Nível	1	2	3	4	5	6	7
1º Critério	**Total**	244	179	75	41	16	12	2
	Alterados	90	77	34	16	6	7	1

(**Fonte:** desenvolvido pelo autor.)

Com base na análise dos resultados, a aplicação deste critério se mostrou bastante eficiente, pois diversos artigos que não eram necessários estavam presentes na estrutura dos SNs e foram excluídos, deixando os SNs mais simples, claros e de fácil compreensão.

4.4 Remoção de pronomes

O critério de remoção de pronomes tem o objetivo de simplificar e normalizar o sintagma nominal sem alterar o significado do SN e preservar a palavra-chave do autor.

A aplicação deste critério também foi utilizada por Lopes (2012), porém as restrições estabelecidas na presente pesquisa foram: não alterar a palavra-chave do autor, não alterar o sentido do SN e não remover pronomes que são o núcleo do sintagma nominal.

Em alguns casos, é possível que o núcleo do SN seja um pronome, adjetivo ou verbo no particípio passado, porém a grande maioria possui como núcleo um substantivo.

Lopes (2012) dividiu a aplicação deste critério em duas etapas. A primeira correspondeu a remoção de pronomes que estavam apenas no início dos SNs, e a segunda etapa realizou a remoção dos demais pronomes. O resultado obtido pela autora mostra que através da primeira etapa houve um ajuste de 12.793 SNs de um total de 189.146 SNs, aproximadamente 6,7% do total. O resultado obtido através da segunda etapa foi de um ajuste de 18.230 SNs de um total de 189.146 SNs, aproximadamente 9,6% do total.

Percebe-se que os resultados obtidos por Lopes (2012) nas duas etapas, através da aplicação da heurística de ajuste de pronomes foram bastante satisfatórios. Portanto, nesta dissertação este critério será utilizado com a adoção das duas etapas, removendo todos os pronomes que não alteram o sentido do SN.

Observando o exemplo a seguir é possível perceber melhor como foi feita a aplicação desse critério. Temos os seguintes SNs: “Deste paradigma” e “Alguns conceitos complementares necessários ao entendimento do assunto”, através do critério de remoção de pronomes, temos a remoção do pronome demonstrativo: “Deste”, e do pronome indefinido: “Alguns”. O SN que resulta através da aplicação desse critério nesses SNs são: “Paradigma” e “Conceitos complementares necessários ao entendimento do assunto”.

Os exemplos apresentados acima podem ser melhor observados no Apêndice D, pois lá é possível ver os demais SNs alterados devido a esse critério.

No Apêndice D estão discriminados cada um dos SNs que tiveram alguma alteração através desse critério, através dele é possível identificar os pronomes que

foram removidos da estrutura do SN e perceber que tanto o SN como as palavras-chave dos autores foram preservados.

Durante a aplicação do critério de remoção de pronomes não houve nenhuma alteração nas palavras-chave e na semântica dos SNs, portanto não teve nenhum problema durante a sua aplicação.

Com a aplicação do critério de remoção de pronomes para a normalização dos SNs, foram feitas alterações em 6 dos 569 SNs, cerca de 1,05% do total. As alterações foram bem balanceadas entre os níveis dos SNs (Quadro 28).

Quadro 28 – Nível dos SNs alterados no 2º critério

	Nível	1	2	3	5
2º Critério	**Total**	244	179	75	16
	Alterados	2	2	1	1

(**Fonte:** desenvolvido pelo autor.)

Através do Quadro 29 é possível visualizar, por documento, o número de SNs a normalizar e quantos sofreram alterações depois da aplicação deste critério de normalização.

Quadro 29 – Número de SNs no 2º Critério da Etapa 2

NÚMERO DE SNs ANTES(SNa) E ALTERADOS(SNd)												
Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**
DOC. 1	3	0	DOC. 16	1	0	DOC. 31	6	0	DOC. 46	26	0
DOC. 2	9	0	DOC. 17	0	0	DOC. 32	15	0	DOC. 47	5	1
DOC. 3	11	0	DOC. 18	3	0	DOC. 33	8	0	DOC. 48	13	0
DOC. 4	13	0	DOC. 19	9	0	DOC. 34	7	0	DOC. 49	0	0
DOC. 5	13	0	DOC. 20	2	0	DOC. 35	33	3	DOC. 50	4	0
DOC. 6	5	0	DOC. 21	1	0	DOC. 36	4	0	DOC. 51	18	0
DOC. 7	10	0	DOC. 22	18	1	DOC. 37	4	1	DOC. 52	7	0
DOC. 8	14	0	DOC. 23	13	0	DOC. 38	20	0	DOC. 53	8	0
DOC. 9	12	0	DOC. 24	24	0	DOC. 39	4	0	DOC. 54	15	0
DOC. 10	3	0	DOC. 25	7	0	DOC. 40	8	0	DOC. 55	10	0
DOC. 11	3	0	DOC. 26	4	0	DOC. 41	18	0	DOC. 56	3	0
DOC. 12	6	0	DOC. 27	12	0	DOC. 42	14	0	DOC. 57	14	0
DOC. 13	1	0	DOC. 28	7	0	DOC. 43	20	0	DOC. 58	14	0
DOC. 14	5	0	DOC. 29	11	0	DOC. 44	10	0	DOC. 59	0	0
DOC. 15	0	0	DOC. 30	15	0	DOC. 45	7	0	DOC. 60	17	0

(**Fonte:** desenvolvido pelo autor.)

Os pronomes identificados nos SNs foram do tipo indefinido, demonstrativo e possessivo, especificamente os pronomes: alguns, muitas, algumas, suas e deste. Todos os pronomes foram devidamente removidos e em todos os casos foram respeitados os limites definidos para a sua aplicação.

No Apêndice D estão disponíveis todos os SNs que tiveram a aplicação do critério de remoção de pronomes, nele é possível observar que os 6 pronomes removidos através da aplicação desse critério não causaram nenhum impacto negativo no SN e também não alterou a palavra-chave presente no SN.

Os resultados obtidos através da aplicação desse critério no *corpus* da pesquisa foram semelhantes aos resultados do trabalho de Lopes (2012). O número de SNs que sofreu alterações relacionadas aos pronomes em sua estrutura foram baixos, mas isso não indica que a aplicação desse critério seja irrelevante.

Mesmo tendo um número inferior ao critério de remoção de artigos, o critério de remoção de pronomes também se demonstra eficaz, pois permitiu remover dos SNs pronomes que não são necessários para manter o sentido do SN.

A aplicação manual desse critério se mostrou bastante simples, pois foi fácil a identificação dos pronomes e em nenhum dos casos havia a violação dos limites definidos para a aplicação desse critério.

Portanto, é possível afirmar que a aplicação automática desse critério é plausível, pois não foram identificados problemas que possam causar dificuldades para um sistema.

4.5 Remoção de advérbios

O critério de remoção de advérbios baseou-se em uma das heurísticas de Lopes (2012), porém a sua aplicação é diferente da forma que a autora realizou.

No trabalho de Lopes (2012), o critério relacionado a advérbios era de recusa total do SN, ou seja, todos os SNs que iniciassem com advérbios eram recusados como conceitos.

Diferente da forma que foi utilizado pela autora, o critério aqui foi aplicado de forma menos rigorosa. Os SNs que apresentassem advérbios em sua estrutura, independente de no início, meio ou fim, teriam apenas esse advérbio removido, caso não alterassem o sentido do SN. O motivo para a não remoção completa do SN é a

possível existência de SNs que, mesmo com advérbios, tenham a capacidade de serem descritores.

Lopes (2012), através da aplicação do critério de advérbios, obteve um total de 650 SNs recusados como conceito de um total de 189.146 SNs, cerca de 0,3%. O resultado pode parecer irrelevante em questão de números, mas, conforme a autora, contribuiu significativamente para que SNs de significado vazio fossem recusados.

Na presente pesquisa, a aplicação do critério de advérbios não apresentou nenhuma alteração aos SNs. A causa desse resultado foi identificada como a aplicação das regras presentes na etapa 1 que também tinham a eliminação de alguns advérbios.

Através da aplicação das regras definidas na etapa 1, foram removidos alguns advérbios que estavam presentes nos SNs. Dessa forma, os SNs que chegaram até essa etapa do método não apresentavam advérbios na sua estrutura.

Ao comparar o resultado apresentado por este critério com o resultado dos dois critérios anteriores, podemos afirmar que a aplicação desse critério não se mostrou necessária para a normalização dos SNs no presente *corpus*. Todavia, é possível que a aplicação desse critério em outro *corpus* possa apresentar um resultado diferente.

Sendo assim podemos considerar que o critério de remoção de advérbios não se faz necessário para a normalização no presente *corpus*, pois não apresentou nenhuma alteração aos SNs que foram submetidos ao processo de normalização.

4.6 Remoção de numerais

Este critério foi desenvolvido com base nos trabalhos de Martins (2014) e Lopes (2012).

No trabalho de Martins (2014), o autor utilizou a remoção de numerais antes do processo de *stemming*, a fim de avaliar a alteração gerada pela remoção, e concluiu que a contribuição deste critério não foi tão significante.

Após concluir que a remoção de numerais não foi satisfatória, Martins (2014) testou o processo de *stemming* nos SNs e obteve uma redução de 30% no corpo do texto, concluindo que o *stemming* contribuiu mais do que a remoção de numerais.

Lopes (2012) utilizou em uma de suas heurísticas o critério de recusa de numerais. Segundo a autora, através da recusa de SNs que apresentassem numerais

de forma escrita ou de caracteres numéricos (dígitos), foi obtida uma recusa de 30.969 SNs de um total de 186.146, aproximadamente 16% dos SNs.

Segundo Lopes (2012) a aplicação do critério de recusa de numerais foi bastante importante, pois permitiu que SNs que não se caracterizavam como conceitos fossem descartados, por exemplo: “três meses” e “ano 2000”.

Baseado nas experiências de Lopes (2012) e Martins (2014), decidiu-se adotar nesta pesquisa o critério de remoção de numerais de forma removesse apenas os numerais presentes nos SNs.

A aplicação desse critério tem o objetivo de eliminar numerais escritos por extenso ou em algarismos (romanos e arábicos).

O resultado obtido através da aplicação do critério de remoção de numerais foi baixo, porém satisfatória, 6 SNs foram alterados de um total de 569 SNs, aproximadamente 1% dos SNs (Quadro 30).

Quadro 30 – Número de SNs no 4º Critério da Etapa 2

NÚMERO DE SNs ANTES(SNa) E ALTERADOS(SNd)												
Documento	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**	**Documento**	**SNa**	**SNd**
DOC. 1	3	0	DOC. 16	1	0	DOC. 31	6	0	DOC. 46	26	0
DOC. 2	9	0	DOC. 17	0	0	DOC. 32	15	0	DOC. 47	5	0
DOC. 3	11	0	DOC. 18	3	0	DOC. 33	8	1	DOC. 48	13	0
DOC. 4	13	0	DOC. 19	9	0	DOC. 34	7	0	DOC. 49	0	0
DOC. 5	13	0	DOC. 20	2	0	DOC. 35	33	1	DOC. 50	4	1
DOC. 6	5	0	DOC. 21	1	0	DOC. 36	4	0	DOC. 51	18	0
DOC. 7	10	0	DOC. 22	18	0	DOC. 37	4	0	DOC. 52	7	0
DOC. 8	14	0	DOC. 23	13	0	DOC. 38	20	0	DOC. 53	8	0
DOC. 9	12	0	DOC. 24	24	0	DOC. 39	4	0	DOC. 54	15	0
DOC. 10	3	0	DOC. 25	7	0	DOC. 40	8	0	DOC. 55	10	0
DOC. 11	3	0	DOC. 26	4	0	DOC. 41	18	3	DOC. 56	3	0
DOC. 12	6	0	DOC. 27	12	0	DOC. 42	14	0	DOC. 57	14	0
DOC. 13	1	0	DOC. 28	7	0	DOC. 43	20	0	DOC. 58	14	0
DOC. 14	5	0	DOC. 29	11	0	DOC. 44	10	0	DOC. 59	0	0
DOC. 15	0	0	DOC. 30	15	0	DOC. 45	7	0	DOC. 60	17	0

(**Fonte:** desenvolvido pelo autor.)

Diferente do critério de remoção de artigos e do critério de remoção de pronomes, percebe-se que a aplicação do critério de remoção de numerais ocorreu apenas nos SNs de nível 1, 2 e 3, conforme o quadro a seguir.

Quadro 31 – Nível dos SNs Alterados Pelo 4º Critério

	Nível	1	2	3
4º Critério	**Total**	244	179	75
	Alterados	3	1	2

(**Fonte:** desenvolvido pelo autor.)

Durante a aplicação desse critério ocorreram duas situações extraordinárias. A primeira foi o SN “1995 e 2000”, que foi completamente excluído da lista de SNs, pois não existiria mais elemento gramatical para representar um SN, apesar do PyPLN considerar essa frase como uma unidade mínima de significado. A segunda foi o SN “545% titulados”, que não se adequava como SNs foi ficava apenas o verbo após a remoção do numeral, logo também foi excluído da lista de SNs.

Devido ao número baixo de SNs alterados devido a essa regra, optou-se por apresentar no quadro 32 os SNs antes e depois da alteração.

Quadro 32 – SNs Alterados Pelo 4º Critério

Doc.	**SN Antes**	**SN Depois**
Art. 33	1995 a 2000	EXCLUIDO
Art. 35	versão 40 do **Html**	versão do **Html**
Art. 41	545 % titulados	EXCLUIDO
Art. 41	dos 5 Programas em **Ciência da Informação**	dos Programas em **Ciência da Informação**
Art. 41	5 Programas em **Ciência da Informação**	Programas em **Ciência da Informação**
Art. 50	no século XXI com	no século com

(**Fonte:** desenvolvido pelo autor.)

No quadro 32 é possível identificar os numerais que foram removidos da estrutura dos SNs e as palavras-chave representadas em negrito e sublinhado. Percebe-se que a aplicação do critério de remoção de numerais não apresentou nenhum problema e não alterou o significado do SN e preservou a palavra-chave.

Observando o quadro 32 percebe-se que dos 6 SNs alterados pelo critério de remoção de numerais, apenas 3 deles continham a palavra-chave do autor em sua estrutura. Entretanto, mesmo SNs que não apresentem a palavra-chave do autor em sua estrutura, também podem ser alterados com a aplicação das regras.

Através dos resultados obtidos através do critério de remoção de numerais, fica claro que a sua aplicação para a normalização de SNs é útil, pois, assim como o

critério de remoção de pronomes, possibilitou que os SNs extraídos pelo PyPLN expressassem melhor seus significados e ficassem mais próximos de um termo normalizado.

Em consonância com os demais critérios, o critério de remoção de numerais não apresentou problemas durante sua aplicação, portanto não são necessários ajustes para a sua utilização de forma automática por um sistema.

4.7 Remoção de verbos

O critério de remoção de verbos foi desenvolvido através da análise dos SNs extraídos do PyPLN antes da aplicação da etapa 1.

O objetivo desse critério foi verificar a existência de verbos desnecessários na estrutura dos SNs e se sua remoção não descaracterizaria os SNs.

A aplicação do critério de remoção de verbos foi feita apenas em verbos que não estivessem no particípio passado, ou seja, os demais verbos que não estivessem no particípio passado poderiam ser excluídos da estrutura do SN.

Através desse critério não houve nenhuma alteração nos SNs, isso é atribuído ao fato de que durante a etapa 1, algumas regras já removiam verbos existentes nos SNs, ou seja, não sobraram verbos para serem removidos durante a etapa 2.

Esse critério apresentou um resultado idêntico ao critério de remoção de advérbios. Portanto, é possível afirmar que a sua aplicação para a normalização dos SNs não se mostra necessária para o *corpus* desta pesquisa. Todavia existe a possibilidade desse critério ser útil para a normalização de outro *corpus*.

4.8 Remoção de preposições e conjunções do início e fim dos SNs

Esse critério foi desenvolvido com base na análise dos SNs que resultaram da extração do PyPLN.

O critério de remoção de preposições e conjunções do início e fim dos SNs foi aplicado de forma semelhante ao critério de remoção de artigos. Através desse critério foi possível eliminar preposições e conjunção como: “de”, “por”, “pelo”, “pela”, ou, etc.

Para compreender melhor a aplicação desse critério é indicado que seja observado o Apêndice E, pois, nele estarão presentes todos os SNs alterados com base nesse critério.

Para exemplificar a aplicação desse critério temos os seguintes SNs: “da universidade”, “Estudos sociais da ciência e tecnologia ou” e “no interior”. Observe que temos a presenta das preposições “da” e “no” e da conjunção “ou”. Através da aplicação do critério, eliminamos esses elementos que estejam localizados no início ou no fim do SN. Os SNs que resultaram através desse critério foram: “Universidade”, “Estudos sociais da ciência e tecnologia” e “Interior”.

Em todos os casos que foram exemplificados a aplicação desse critério não apresentou problemas, pois foi respeitado o significado do SN e a palavra-chave.

No total, 110 SNs (19%) foram alterados através desse critério. Um resultado bastante expressivo, assim como o resultado apresentado pelo critério de remoção de artigos.

Dos 110 SNs, 2 possuiam a conjunção “ou” no fim do SN. Os demais SNs tiveram as preposições “da”, “do”, “das”, “dos”, “na”, “no”, “nos” e “pelos”, removidas de onde estavam posicionadas no SN, portanto, todos os removidos se encontravam no início do SN.

O número de SNs alterados por documento através do critério de remoção de pronomes pode ser observado no quadro 33.

Quadro 33 – Número de SNs no 6º Critério da Etapa 2

NÚMERO DE SNs ANTES(SNa) E ALTERADOS(SNd)												
Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**	Documento	**SNa**	**SNd**
DOC. 1	3	0	DOC. 16	1	0	DOC. 31	6	1	DOC. 46	26	5
DOC. 2	9	0	DOC. 17	0	0	DOC. 32	15	5	DOC. 47	5	0
DOC. 3	11	0	DOC. 18	3	3	DOC. 33	8	2	DOC. 48	13	6
DOC. 4	13	0	DOC. 19	9	0	DOC. 34	7	1	DOC. 49	0	0
DOC. 5	13	0	DOC. 20	2	0	DOC. 35	33	13	DOC. 50	4	1
DOC. 6	5	0	DOC. 21	1	1	DOC. 36	4	2	DOC. 51	18	4
DOC. 7	10	0	DOC. 22	18	4	DOC. 37	4	0	DOC. 52	7	3
DOC. 8	14	0	DOC. 23	13	2	DOC. 38	20	0	DOC. 53	8	3
DOC. 9	12	2	DOC. 24	24	5	DOC. 39	4	1	DOC. 54	15	1
DOC. 10	3	1	DOC. 25	7	2	DOC. 40	8	0	DOC. 55	10	3
DOC. 11	3	0	DOC. 26	4	0	DOC. 41	18	4	DOC. 56	3	0
DOC. 12	6	1	DOC. 27	12	0	DOC. 42	14	5	DOC. 57	14	3
DOC. 13	1	0	DOC. 28	7	2	DOC. 43	20	3	DOC. 58	14	3
DOC. 14	5	0	DOC. 29	11	2	DOC. 44	10	5	DOC. 59	0	0
DOC. 15	0	0	DOC. 30	15	6	DOC. 45	7	2	DOC. 60	17	3

(**Fonte:** desenvolvido pelo autor.)

Assim como o critério de remoção de artigos, o critério de remoção de preposições e conjunções apresentou um grande número de alterações nos SNs de nível 1 e 2 e poucas alterações nos SNs de nível 3, 4 e 5 (quadro 34.

Quadro 34 – Nível dos SNs Alterados Pelo 6º Critério

	Nível	1	2	3	4	5
6º Critério	**Total**	244	179	75	41	16
	Alterados	56	32	13	6	3

(**Fonte:** desenvolvido pelo autor.)

Com base nesses resultados, é possível afirmar que o critério de remoção de preposições e conjunções se mostrou bastante eficiente, pois teve um resultado bastante expressivo sem alterar o significado do SN e preservando a palavra-chave.

No que tange à aplicação manual desse critério, não foram encontradas dificuldades, pois as preposições e conjunções que foram removidas do início e do fim dos SNs eram desnecessárias e a sua remoção não causou nenhum problema para a estrutura e para as palavras-chave.

Sobre a aplicação automática desse critério para a normalização dos SNs, é possível afirmar que a sua aplicação é viável por um sistema.

4.9 Remoção de sufixos

Antes de apresentar os resultados do critério de remoção de sufixo, e a fim de facilitar a compreensão dos valores obtidos através dos seis critérios da etapa 2, é possível observar, no quadro 35, abaixo, a quantidade de SNs que sofreram alterações em cada um dos seis critérios.

Quadro 35 – Nível dos SNs Alterados Pelos Critérios da Etapa 2

	Nível	1	2	3	4	5	6	7	Total
1º Critério	**Alterados**	90	77	34	16	6	7	1	231
2º Critério	**Alterados**	2	2	1	0	1	0	0	6
3º Critério	**Alterados**	0	0	0	0	0	0	0	0
4º Critério	**Alterados**	3	1	2	0	0	0	0	6
5º Critério	**Alterados**	0	0	0	0	0	0	0	0
6º Critério	**Alterados**	56	32	13	6	3	0	0	110

(**Fonte:** desenvolvido pelo autor.)

Ao analisar o quadro 35 é evidente que os SNs que mais sofreram alterações foram os de nível 1 e 2. O principal motivo para isto é que grande parte dos SNs extraídos através do PyPLN são destes níveis (Quadro 36).

Quadro 36 – Total de SNs Alterados Por Nível

Nível	1	2	3	4	5	6	7
Total de SNs	151	112	50	22	10	7	1

(**Fonte:** desenvolvido pelo autor.)

Analisando o número total de SNs, constatou-se que cerca de 62% dos SNs (353 de 569) tiveram alguma alteração em sua estrutura como consequência da aplicação de um ou mais critérios utilizados para a normalização.

De acordo com o Quadro 37, é possível constatar os seguintes dados:

- Os SNs de nível 1 tiveram 151 alterações de um total de 244 SNs, cerca de 61,8%;
- Os SNs de nível 2 tiveram 112 alterações de um total de 179, cerca de 62,5%;
- Os SNs de nível 3 tiveram 50 alterações de um total de 75 SNs, cerca de 66%;
- Os SNs de nível 4 tiveram 22 alterações de um total 41 SNs, ou seja, 53%;
- Os SNs de nível 5 tiveram 10 alterações de um total de 16 SNs, cerca de 62,5%;
- Os SNs de nível 6 tiveram 7 alterações de um total de 12 SNs, cerca de 85,7%;
- Os SNs de nível 7 tiveram 1 alterações de um total de 2, exatamente 50%;

Quadro 37 – Total de SNs Por Nível e Total de SNs Alterados Por Nível

Nível	1	2	3	4	5	6	7
Total de SNs	244	179	75	41	16	12	2
SNs Alterados	151	112	51	23	12	7	1

(**Fonte:** desenvolvido pelo autor.)

Todavia, o número de SNs que possuíam a palavra-chave em sua estrutura e que tiveram a aplicação de algum dos critérios uma vez ou mais em sua estrutura foi de 181 SNs.

Em todos esses 181 SNs não houve a alteração nem no significado do SN e nem da palavra-chave, ou seja, os critérios utilizados para a normalização dos SNs

não apresentaram problemas e, portanto, podem ser utilizados para a normalização de SNs.

No total, 221 SNs não tiveram alterações através dos critérios aqui estabelecidos. Desses 220 SNs, 189 possuem as palavras-chave dos autores em sua estrutura, enquanto que 32 não possuem.

O critério de remoção de sufixos, último a ser aplicado, tem a função de verificar se as alterações feitas através da aplicação dos critérios anteriores atingiram o objetivo da proposta de normalização.

A proposta de normalização foi avaliada através da verificação da existência de termos canonizados do Tesauro Brasileiro de Ciência da Informação na estrutura dos SNs que passaram pelos critérios de normalização.

Este tesauro é uma lista organizada por profissionais da área e é considerada, por diversos autores, uma ferramenta extremamente importante para o auxílio do profissional indexador. Além disso, o *corpus* trabalhado na presente pesquisa pertence ao mesmo campo científico abrangido pelo tesauro.

Os SNs que passaram pelo processo de remoção de sufixo não foram apenas os que tiveram alterações realizadas pelos outros critérios, mas sim todos. Esta escolha possibilitou verificar se os SNs, alterados ou não, podem ser considerados SNs normalizados, ou seja, qualificados como descritores dos documentos.

O processo de remoção de sufixo foi bastante simples. Os SNs foram selecionados e transferidos para o *software* Luke, que, por sua vez, automaticamente, retornava uma lista com todos os SNs com as palavras sem seus respectivos sufixos.

A fim de verificar se os SNs podem ser considerados normalizados e qualificados como descritores, foram submetidos ao processo de remoção de sufixo 567 SNs. Destes, 375 SNs apresentaram no mínimo 1 termo do TBCI em sua estrutura, ou seja, são considerados normalizados e isto equivale a aproximadamente 66,4% dos SNs.

A quantidade de termos do TBCI que foram encontrados em cada um dos SNs pode ser observada na coluna “"Idênticos", "Não São Termos do TBCI" e "Contêm Termos do TBCI"” do Apêndice F. Neste Apêndice estão descritos todos os SNs que passaram por todos os critérios, inclusive o critério de remoção de sufixo, nele também é apresentado todo o processo de remoção de sufixo e a quantidade de termos do TBCI presente nos SNs.

Com a aplicação do critério de remoção de sufixo foi possível identificar a presença dos termos do TBCI nos SNs.

Com relação aos SNs idênticos, foi constatado que de 567 SNs, 179 SNs eram idênticos.

No total, 192 SNs não foram considerados descritores, pois não estavam presentes no tesauro.

Ao analisar todos os SNs considerados normalizados com base nos SNs que apresentavam algum termo do TBCI (Quadro 38), é possível afirmar que: todos os SNs de nível 6 e 7 foram considerados normalizados, indicando um resultado de 100%; os SNs de nível 1 tiveram 49% dos SNs considerados normalizados; os SNs de nível 2 tiveram 77% dos SNs considerados normalizados; e os SNs de nível 3 tiveram 80% dos SNs considerados normalizados.

Quadro 38 – Total de SNs Normalizados

Nível	1	2	3	4	5	6	7
Total de SNs	244	179	75	41	16	12	2
SNs Normalizados	121	138	60	31	13	12	2

(**Fonte:** desenvolvido pelo autor.)

Ao analisar manualmente os SNs que foram considerados normalizados, ficou claro que todos os SNs que passaram pelos critérios de normalização expressavam valor descritivo para o documento e, portanto, poderiam ser utilizados como descritores em bases de dados que utilizem da indexação automática por SNs.

Para facilitar a compreensão dos SNs normalizados, temos os exemplos dos seguintes SNs: “Gestão do conhecimento”, “Hipertexto” e “Ação efetiva”. Os SNs “Gestão do conhecimento” e “Ação efetiva” tiveram alterações devido aos critérios estabelecidos para a etapa 2, entretanto o SN “Hipertexto” não teve alterações, porém através da aplicação do critério de sufixo foi constatado a presença dos SNs “Gestão do conhecimento” e “Hipertexto” no TBCI, portanto ambos foram considerados descritores. Por outro lado, o SN “Ação efetiva” não se encontrava o seu sufixo no TBCI, logo não foi possível classifica-lo como um SN normalizado.

A aplicação automática do critério de remoção de sufixo para a verificação da normalização de SNs é possível, pois trata-se de um processo simples onde os SNs tem o seu sufixo removido temporariamente apenas para a comparação com os termos presentes em algum vocabulário controlado.

Por fim, para facilitar a compreensão de os critérios aplicados na segunda etapa está disponível no Apêndice F alguns exemplos da aplicação desses critérios nos SNs. Devido ao grande número de SNs e a extensa planilha que foi desenvolvida com os dados dessa pesquisa, não foi possível disponibilizar de forma física todos os SNs alterados, portanto, no Apêndice F também está disponível um *link* que disponibilizará em formato eletrônico a planilha com todos os SNs.

No quadro 39 é apresentado os valores dos SNs alterados ou não, que possuem ou não as palavras-chave dos autores, em relação ao TBCI. São apresentados: o número desses SNs que são idênticos aos termos do TBCI; os que contemplam termos do TBCI e os que não contém termo do TBCI.

Quadro 39 – Quadro de relação dos SNs e TBCI

	Idênticos a termos do TBCI	**Contemplam termos do TBCI**	**Não contém termo do TBCI**	**TOTAL**
SNs finais com palavras-chave alterados	95	117	66	251
SNs finais sem palavras-chave alterados	12	7	50	69
SNs finais com palavras-chave não alterados	68	70	50	188
SNs finais sem palavras-chave não alterados	1	5	26	32
TOTAL	176	199	192	567

(**Fonte:** desenvolvido pelo autor.)

No quadro 39 é observado que o total de SNs sem palavras-chave que não contem termo do TBCI é de 50, entretanto, levando em conta que o SN “1995 a 2000” e o SN “545% titulados” foram excluídos com a aplicação do critério de remoção de numerais e por não se qualificarem mais como um SN, o número de SNs sem palavras-chave alterados seria 48.

5 CONCLUSÃO

A principal motivação para o desenvolvimento da pesquisa foi a grande importância que a indexação possui para a recuperação da informação e a atual dificuldade enfrentada pelos profissionais indexadores ao tentar indexar manualmente um grande fluxo informacional.

O objetivo da pesquisa foi desenvolver uma proposta de normalização de SNs extraídos automaticamente em termos canônicos. Esta proposta foi composta por critérios fundamentados na CI e levou em consideração os princípios e conceitos da linguagem documentária como principal base científica.

A presente pesquisa conseguiu alcançar todos os objetivos gerais e específicos que foram estabelecidos.

Com o propósito de atingir o objetivo da pesquisa, foi necessário o delineamento de uma metodologia que possibilitasse normalizar os SNs extraídos do *corpus* da pesquisa. Sendo assim, aqui são apresentadas as considerações finais sobre a metodologia utilizada e o método proposto para a normalização dos SNs extraídos automaticamente.

Uma etapa importante da metodologia, mas que não está diretamente ligada ao processo de normalização, e sim ao processo de seleção dos SNs, foi o teste de revocação.

Através do teste de revocação foi possível descartar diversos SNs que não contribuíam para a descrição dos documentos, ou seja, não tinham condições de serem considerados descritores. O teste de revocação possibilitou uma amostragem mais adequada para os SNs que seriam utilizados na proposta de normalização.

Sobre a aplicação das etapas e dos critérios propostos para a normalização dos SNs, a principal regra limitadora das alterações seria a preservação do sentido expressado pelos SNs e não alterar a palavra-chave do autor.

A etapa 1 se mostrou bastante positiva, pois conseguiu diminuir o número de SNs extensos, criando novos SNs de níveis inferiores e mais adequados como termos descritores dos documentos.

Através da aplicação das regras desenvolvidas para etapa 1, é evidente que sua utilização para um sistema automático de normalização dos SNs é positiva, pois as regras conseguiram, sem falha, desenvolver novos SNs descritores para os documentos.

O critério de remoção de artigos do início e fim dos SNs foi considerado o mais importante de toda a proposta de normalização, pois foi o responsável por um grande número de alterações nos SNs, tornando-os mais claros e objetivos.

O critério de remoção de pronomes foi um dos que menos proporcionou alterações na estrutura dos SNs. O principal motivo para isto foi o baixo índice de pronomes presentes na estrutura dos SNs selecionados para a normalização.

A principal preocupação durante a aplicação deste critério foi o mesmo citado por Lopes (2012), onde existia a possibilidade de SNs apresentarem, como núcleo, determinados pronomes e, sendo assim, não deveriam ser excluídos do SN. Portanto, este critério ficou limitado na presente pesquisa a satisfazer a três regras: preservar o sentido do SN, não alterar a palavra-chave do autor e não excluir os pronomes que fossem núcleo dos SNs. Entretanto, não houve nenhum SN que apresentasse essas características.

Com base nestas regras, o critério de remoção de pronomes conseguiu atingir seu objetivo de apresentar, de forma mais clara, os SNs.

Apesar deste critério apresentar um número baixo de SNs alterados, ele teve um papel fundamental para a normalização dos SNs, pois para que os SNs fossem considerados normalizados era preciso que em sua estrutura não existissem elementos que não contribuíssem para a descrição. Portanto, é possível afirmar que a aplicação deste critério é eficiente para a normalização dos SNs.

A aplicação desse critério através de um sistema automático é possível, pois durante a sua aplicação manual não se percebeu nenhuma dificuldade que pudesse atrapalhar um sistema de utiliza-lo de forma irrestrita.

O critério de remoção de advérbios foi um dos que não apresentaram alterações na estrutura dos SNs. Esse critério foi desenvolvido com base no critério similar utilizado por Lopes (2012).

Na pesquisa de Lopes (2012), a autora apresenta um resultado positivo através da aplicação da heurística de recusa de SNs que iniciam com advérbio.

Na presente pesquisa verificou-se que o critério de remoção de advérbios presentes na estrutura dos SNs não é necessário para a normalização dos SNs, pois não trouxe benefício algum para a normalização dos SNs, isso se deve ao fato de que muitos desses advérbios foram removidos durante a etapa 1.

O quarto critério utilizado para alterar os SNs e assim aproximá-lo ou torná-lo um SN normalizado, foi o critério de remoção de numerais. Igualmente ao critério de

remoção de artigos, acreditava-se que a utilização deste critério seria interessante para a normalização dos SNs, e isto se confirmou, como descrito na análise dos resultados.

O critério de remoção de numerais conseguiu, de forma bastante precisa, atingir os objetivos que lhe foram impostos, como: aproximar o SN de um SN normalizado; preservar o significado da frase; e descartar elementos desnecessários para a estrutura dos SNs.

Ainda através deste critério foi possível adequar os SNs à diretriz utilizada no processo de criação de um vocabulário controlado: a diretriz de exclusão de numerais (escritos e dígitos). Estas adequações aproximaram os SNs dos termos considerados descritores, presentes no vocabulário controlado.

Quanto a aplicação automática do critério de remoção de numerais, verificou-se que pode ser feita sem nenhuma dificuldade.

O critério de remoção de verbos não apresentou alteração nos SNs durante a etapa 2, isso se deve ao fato de que esses verbos foram removidos durante a etapa 1 e não houve a permanência de nenhum outro verbo que não estivesse no particípio passado.

O último critério responsável por alterações na estrutura dos SNs, antes da aplicação do critério de remoção de sufixo (responsável pela verificação da normalização), foi o critério de remoção de preposições e conjunções do início e fim dos SNs.

O resultado obtido pelo critério de remoção de preposições e conjunções foi bastante similar ao critério de remoção de artigos. Ele foi considerado o segundo critério responsável pelo maior número SNs alterados. Esse critério foi bastante satisfatório pois preservou o sentido expressado pelos SNs e não alterou a palavra-chave. A aplicação desse critério conseguiu apresentar SNs mais próximos dos termos canonizados.

A aplicação automática do critério de remoção de preposições e conjunções do início e fim dos SNs é favorável, pois seus resultados demonstraram um grande número de alterações dos SNs e não houve dificuldade na sua aplicação.

Ao analisar os critérios de normalização da etapa 2, conclui-se que a aplicação dos critérios de: remoção de artigos; remoção de pronomes; remoção de numerais; e remoção de preposições e conjunções, foram importantes para o processo de normalização dos SNs. Todavia, o critério de remoção de artigos e o critério de

remoção de preposições e conjunções foram os mais importantes, pois apresentaram um maior índice de alterações.

Sendo assim é possível determinar que os critérios de remoção de advérbios e verbos, não contribuíram em nenhum aspecto para a normalização dos SNs no presente *corpus*. Todavia, a aplicação desses critérios em outro *corpus* pode apresentar resultados diferentes.

Através desta pesquisa foi possível confirmar as afirmativas propostas por Lopes (2012) sobre a importância dos critérios referentes aos artigos e pronomes.

Após a aplicação dos seis critérios, foi utilizado o critério de remoção de sufixos, responsável por permitir o casamento entre SNs e termos do TBCI e a posterior classificação dos SNs como SNs normalizados. Os resultados apresentados por este critério foram bastante satisfatórios e indicaram que os critérios utilizados para a proposta de normalização conseguiram normalizar parte dos SNs, como também identificou que maioria dos SNs não alterados também se encontravam normalizados.

Com essa pesquisa, acredita-se que será possível contribuir para a CI com informações relativas a indexação automática por SNs, pois com o método normalização proposto e os dados apresentados e discutidos nessa pesquisa é possível que demais pesquisadores possam desenvolver novas técnicas relativas a indexação automática por SNs.

Acredita-se que com os resultados dessa pesquisa será possível aprofundar as pesquisas sobre a normalização dos SNs e também aplicar essa normalização em sistemas de indexação por SNs.

Com relação as indagações levantadas na problemática da pesquisa é possível respondê-las da seguinte forma:

- Como selecionar os SNs mais relevantes para serem normalizados? No presente *corpus* o método utilizado foi o de teste de revocação, porém em outro *corpus* é possível desenvolver outros métodos que possibilitem selecionar os SNs relevantes.
- Todos os SNs normalizados são possíveis descritores?
  Em parte, pois carregam o valor descritivo, semântico e sintático presente nos SNs, entretanto alguns SNs vazios de significado permaneceram.
- A indexação automática com SNs normalizados descreve corretamente os documentos?
  Sim, pois através dos SNs normalizados, foi possível eliminar elementos gramaticais desnecessários para a descrição dos documentos.

- Os critérios de normalização podem ser usados para detectar e eliminar SNs vazios e SNs irrelevantes?
  Em parte é possível, pois como visto no critério de remoção de numerais, dois SNs foram eliminados por não serem SNs.

É sabido que nem todos os campos científicos possuem um tesauro relativamente atualizado, como o que foi utilizado na presente pesquisa. Entretanto, é possível que esta ferramenta seja substituída por outra e possibilite a mesma avaliação para outras áreas de conhecimento.

Ficou evidente durante a avaliação da proposta de normalização que os SNs ficaram mais claros e objetivos e preservaram sua principal característica que é o sentido do SN.

As limitações encontradas na presente pesquisa foram relativas ao *corpus* e a ferramenta PyPLN. A limitação relativa ao *corpus* foi que os critérios foram desenvolvidos com base no *corpus* dessa pesquisa, ou seja, possivelmente não sendo exaustivo para outro *corpus*. A limitação referente a ferramenta PyPLN foi a instabilidade que a ferramenta apresentou durante a sua utilização, em determinados momentos ela não se encontrava disponível.

A fim de contribuir para que mais pesquisas sejam desenvolvidas, ficam propostas aqui ideias que surgiram durante o desenvolvimento desta pesquisa e que podem ser continuadas em trabalhos futuros:

1. Rever a aplicação dos critérios de remoção de advérbios e verbos em outro corpus, a fim de confirmar a sua eliminação da etapa 2 de normalização;
2. Utilização de *Machine Learning* para a indexação automática por SNs, buscando verificar se é possível melhorar a normalização ou até mesmo a indexação automática por SNs;
3. Mensurar os índices de revocação e precisão antes e depois da etapa de normalização dos SNs;
4. Analisar a normalização de SNs de corpus de outras áreas do conhecimento
5. Analisar a aplicação dos critérios de normalização a todos os SNs, necessitando para isto de uma ferramenta automatizada de normalização de SNs.

O desenvolvimento deste trabalho só foi possível por causa das diversas pesquisas que contribuíram para indexação automática por SNs. Desta forma, espera-se que esta pesquisa sirva como apoio para diversos pesquisadores que desejem dar continuidade a este tema.

REFERÊNCIAS

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais**. Enc. Bibli. R. electrónica de Bibl. Ci. Inform.**, Florianópolis, n.15, 2003.

ANSI/NISO Z39.19: guidelines for the construction, format, and management of monolingual controlled vocabularies. Bethesda, Maryland: NISSO Press, 2005.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676: métodos para análise de documentos: determinação de seus assuntos e seleção de termos de indexação: procedimento**. Rio de Janeiro, 1992.

BAEZA-YARES; RIBEIRO NETO, B. **Modern Information Retrieval**. [S.l]: ACM Press, 1999.

BARBOSA, Ricardo R. Perspectivas profissionais e educacionais em biblioteconomia e ciência da informação. **Ci. Inf**., v. 27, n.1, p. 53-60, jan./abr. 1998.

BARRETO, Aldo de Albuquerque. A condição da informação. **Revista São Paulo em Perspectiva**, v. 16, n. 3, p. 67-74, 2002. Disponível em: Acesso em: 21 nov. 2012

_____. Uma história da Ciência da Informação. In: TOUTAIN, Lídia Maria Batista Brandão (Org.). **Para entender a Ciência da Informação**. Salvador, EDUFBA, 2007, p. 13-34.

BASTOS, S. B. **Análise comparativa entre indexação automática e manual da literatura Brasileira de ciência da informação**. Mestrado, Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 1984.

BECHARA, E. **Moderna gramática portuguesa**. 37. ed. Rio de Janeiro: Nova Fronteira, 2009.

BERNIER, Charles L. Indexing process evaluation. **American Documentation**, v. 16, n. 4, p. 323-328, Out. 1965.

BICK, E. **The Parsing System Palavras**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 505 f. Tese (Doutorado em Linguística), Department Of Linguistics, University Of Aarhus, Arthus, 2000.

BOCCATO, V. R. C. **Avaliação do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias**: um estudo sociocognitivo com protocolo

verbal. Marília, 2009. 301f. Tese (Doutorado e Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista Júlio de Mesquita Filho

BORGES, G.S.B. **Indexação automática de documentos textuais:** proposta de critérios essenciais. 2009. 113 f. Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Minas Gerais, 2009.

BORGES, Graciane S. B.; MACULAN, Benildes C. M. S.; LIMA, Gercina Â. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Inf. Soc.: Est**., João Pessoa, v. 18, n. 2, p. 181-193, maio/ago. 2008.

BORKO, H. Information science: what is it? **American Documentation**, Chicago, v.19, n.1, p.3-5, Jan. 1968.

BRANDT, Mariana; MEDEIROS, Marisa Bräscher Basílio. Folksonomia: esquema de representação do conhecimento?. **TransInformação**, Campinas, v. 22, n. 2, p. 111-121, maio/ago. 2010. Disponível em: <https://www.puc-campinas.edu.br/periodicocientifico/>. Acesso em: 15 ago. 2016

BRASCHER, Marisa; CAFÉ, Lígia. Organização da informação ou organização do conhecimento?. In: ENCONTRO NACIONAL DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9., 2008, São Paulo. **Anais**… São Paulo: USP, 2008.

BURKE, Peter. **Uma história social do conhecimento**: de Gutemberg a Diderot. Rio de Janeiro: Jorge Zahar, 2003. 241 p.

CÂMARA JÚNIOR, A. T. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. Brasília, DF, 2007. 142 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília. Brasília, DF, 2007.

CESARINO, M. A.; PINTO, M. C. M. F. Cabeçalhos de assunto como linguagem de indexação. **Revista da Escola de Biblioteconomia UFMG**, Belo Horizonte, v. 7, n. 2, p. 268-288, set. 1978.

_____. Análise de assunto. **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 1, p. 32-43, jan./jun. 1980

CHAUMIER, J. **Analisis y lenguajes documentales**. Barcelona: Mitre, 1986

CHAUMIER, J. **Analyse et langages documentaires**: Le traitement linguistique de l´information documentaire. Paris. Moderne d´Edition, 1982.

CHIZZOTTI, Antonio. **Pesquisa em ciências humanas e sociais**. 2. ed. São Paulo: Cortez, 1995.

CHU, C. M.; O'BRIEN, A. Subject analysis: the critical first stage in indexing. **Journal of Information Science**, London, v. 19, n. 6 p. 439-54, 1993.

CINTRA, A. M. M. *et al*. **Para entender as linguagens documentárias**. 2. ed. rev. e ampl. São Paulo: Polis, 2002. 92 p.

CLEVELAND, D. B.; CLEVELAND, A. D. **Introduction to indexing and abstracting**. Englewood: Libraries Unlimited, 1990.

COATES, E. J. **Subject catalogues**: headings and structure. London: The Library Association, 1988.

CORRÊA, et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**, Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011.

COYAUD, M. **Introduction a l'etude des langages documentaries**. Paris: Librarie C. Klincksieckm 1966.

DENZIN, N. K. e LINCOLN, Y. S. Introdução: a disciplina e a prática da pesquisa qualitativa. In: DENZIN, N. K. e LINCOLN, Y. S. (Orgs.). **O planejamento da pesquisa qualitativa: teorias e abordagens**. 2. ed. Porto Alegre: Artmed, 2006. p. 15-41

DIAS, Eduardo Wense. Organização do conhecimento no contexto de bibliotecas tradicionais e digitais. In.: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Orgs.). **Organização da informação: princípios e tendências**. Brasília: Briquet de Lemos, 2006. p. 62-75.

DIAS, Eduardo Wense; NAVES, Madalena Martins Lopes. **Análise de Assunto.** Brasília: Thesaurus, 2007.

EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM**, v. 16, n. 2, p. 264- 285, Apr. 1969.

FARROW, J. All in the mind: concept analysis in indexing. **The indexer**, v.19, n.4, Out. 1995. p. 243-247.

FEITOSA, Ailton. **Organização da informação na web**: das tags à web semântica. Brasília: Thesaurus, 2006. 131 p.

FERNEDA, E. **Recuperação da Informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 147 f. Tese (Doutorado em Ciência da Comunicação) – Universidade de São Paulo, São Paulo, 2003.

FIDEL, R. The user-centered approach: how we got there. In: WHEELER, W. J. (ed.). **Saving the user´s time through subject access innovation**. Champaign, The Board of Trustees of The University of Illinois, 2000.

FOGL, J. Relations of the concepts 'information' and 'knowledge'. **International Fórum on Information and Documentation**, The Hague, v.4, n.1, p. 21-24, 1979.

FUJITA, M. S. L. **O contexto da leitura documentária de indexadores de bibliotecas universitárias em perspectiva sócio-cognitiva para a investigação de estratégias de ensino**. Marília: UNESP; CNPq, 2007.

FUJITA, Mariângela S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital Biblioteconomia e Ciência da Informação.**, Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003.

FUJITA, Mariângela Spotti Lopes (Org.). **A indexação de livros**: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais. São Paulo: Cultura Acadêmica, 2009

FUJITA, Mariângela Spotti Lopes; RUBI, Milena Polsinelli. O ensino de procedimentos de política de indexação na perspectiva do conhecimento organizacional: uma proposta de programa para a educação à distância do bibliotecário. **Perspectiva Ciência da Informação.**, Belo Horizonte, v. 11, n. 1, p. 48-67.

GARDIN, J. C. Elemtents d' um modele pour la description des lexiques documentaires. **Bulletin des Biblioteques de France**. n.5, p. 171-182. 1966.

GARVIN, P. L. et al. **Some opinions concerning linguistics and reformation processing.** Washington, D. C.: Center for Applied Linguistics, Maio, 1969.

GIL LEIVA, I. **La automatización de la indización, propuesta teórico-metodológica**: aplicación al área de biblioteconomía y documentación. 1997. 268f. Tese – Universidade de Murcia, Murcia, España, 1997.

GIL LEIVA, Isidoro. **Manual de indización**. Gijón: Trea, 2008.

GIL LEIVA, Isidoro.; FUJITA, Mariângela Spotti Lopes (org). **Política de Indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012.

GIL LEIVA, Isidoro; RUBI, Milena Polsinelli; FUJITA, Mariângela Spotti Lopes. Consistência na indexação em bibliotecas universitárias brasileiras. **Transinformação**, Campinas, v. 20, n. 3, p. 233-253, set./dez., 2008.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010. 185 p. ISBN 9788522458233 (broch.)

GONSALVES, Elisa Pereira. **Conversas sobre iniciação à pesquisa científica**. 4.ed. revisada e ampliada. Campinas, SP: Alínea, 2007.

GONZALEZ, M.; LIMA, VERA L. S. Recuperação de informação e processamento da linguagem natural. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais da Jornada de Mini-Cursos de Inteligência Artificial**, 3. Campinas: [s.n.], v. 3, p.347-395.

GUEDES, Vânia 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, set./dez. 1994.

GUIMARÃES, J. A. C; SALES, R.; GRÁCIO, M. C. C. A dimensão interdisciplinar da análise documental nos contextos brasileiro e espanhol no âmbito da organização do conhecimento. **Datagramazero**, v. 13, n. 6, 2012.

GUIMARÃES, Jose Augusto Chaves. **Abordagens teóricas de tratamento temático da informação (T.T.I.)**: catalogação de assunto, indexação e análise documental. Ibersid: [s.l.], 2009.

GUINCHAT, Claire; MENOU, Michel. **Introdução geral às ciências e técnicas da informação e documentação**. Brasília: IBICT, 1994.

HARTER, Stephen P. Psychological relevance and information science. **Journal of the American Society for Information Science**, v. 43, n. 9, p. 602-6l5, Out. 1992

HJØRLAND, Birger. **Automatic Indexing**. Lifeboat for Knowledge Organization. [s.l.]:[s.n.], 2008. Disponível em: <http://www.iva.dk/bh/lifeboat_ko/CONCEPTS/automatic_indexing.htm >. Acesso em: 15 ago. 2016.

HLAVA, M. M. Automatic indexing: a matter of degree. **Bulletin of the American Society for Information Science and Technology**, v. 29, n. 1, p. 12 – 15, out./nov. 2002.

HORNER, John. **Criticism and evaluation of conventional subject headings lists**. London: Association of Assistant Librarians, 1970.

HUTCHINS, W. J. **Languages of indexing and classification**. Herts: Peter Peregrinus, 1975.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **Guidelines for the establishment and development of multilingual thesauri**: 5964. Londres: BSI, 1985.

_____. **British standard guide to establishment and development of monolingual thesauri**: 2788. Londres: BSI, 1986.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ci. Inf**., Brasília, v.25, n.2, p. 1-18, 1995.

_____. 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, jul./dez. 1997.

_____. Sintagmas Nominais: uma nova proposta para a recuperação de informação. **DataGramaZero** Revista de Ciência da Informação. v. 3, n. 1, fev. 2002.

_____. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ci. Inf**., v. 25, n. 2, p. 1- 18, 1995.

LANCASTER, F. W. **Indexação e Resumos**: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. 2. ed. revista e atualizada. Brasília, DF: Briquet de Lemos, 2004.

LANCASTER, F. W. I**nformation retrieval systems**: characteristics, testing and evaluation. New York: John Wiley & Sons, 1968.

LE COADIC, Yves-François. **A Ciência da Informação**. Brasília: Briquet de Lemos/Livros, 119 p. 1996.

LE GOFF, Jacques. **História e Memória**. Campinas: UNICAMP, 1990.

LE GUERN, Michel. Un analyseur morpho-syntaxique pour l'indexation automatique. **Le Français Moderne**. v. 59, n. 1, p. 22-35, juin 1991.

LOEHRLEIN, Aaron, et. al. A hybrid approach to faceted classification based on analysis of descriptor suffixes. In: Grove, Andrew (Ed.). **ANNUAL MEETING OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**, 68. 2005, Charlotte, US. Proceedings… Charlotte, US: ASIST, 2005. v. 42, p. 1-25.

LOPES, Eunice de Faria. Avaliação de serviços de indexação e resumo: critérios, medidas e metodologia. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, n. 14, v. 2, p. 242-256, set. 1985. Disponível em: <http://www.brapci.ufpr.br/documento.php%3Fdd0%3D0000002652%26dd1%3D5d7bf+&cd=1&hl=pt-BR&ct=clnk&gl=br> Acesso em: 15 ago. 2016.

LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ci. Inf**., Brasília, v. 31, n. 1, p. 41-52, jan. 2002. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100005&lng=en&nrm=iso>. Acesso em: 15 ago. 2016.

LOPES, Lucelene. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012, 156 f. Tese (Doutorado em Ciência da Computação).

Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

MAIA, Luiz Cláudio Gomes.; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da informação**, v. 15, n. 1, p. 154-172, jan./abr. 2010.

MAIA, Luiz Cláudio Gomes. **Uso de Sintagmas Nominais na classificação automática de documentos eletrônicos**. 2008, 158 f. Tese (Doutorado em Ciência da Informação). – Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, 2008.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 6. ed. São Paulo: Atlas, 2009

MARQUES, Otacílio Guedes. **Informação histórica**: recuperação e divulgação da memória do poder judiciário brasileiro. 2007. 133 f. Dissertação (Pós Graduação) - Universidade de Brasília, Brasília, 2007

MARTINS, Agnaldo Lopes. **O uso do sintagma nominal na recuperação de documentos [manuscrito]**: proposta de um mecanismo automático para classificação temática de textos digitais. 2014, 192 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, 2014.

MARTINS, Agnaldo Lopes; SOUZA, Renato Rocha; MELLO, Heliana Ribeiro de. The use of noun phrases in information retrieval: proposing a mechanism for automatic classification. **Knowledge organization in the 21st century**, p. 320-326, 2014. Disponível em: <http://150.164.98.236:8080/ECI/documentos-arquivos/ISKO2014AgnaldoRenatoHelianaFinal.pdf> Acesso em: 01 Jan. 2017.

MEADOW, Charles T.; BOYCE, Bert R.; KRAFT, Donald H. **Text Information Retrieval Systems**. 2. ed. San Diego, CA: Academic Press, 2000.

MEDEIROS, Graziela Martins de. **Organização da informação em repositórios digitais**: implicações do auto-arquivamento na representação da informação. 2010, 274 f. Dissertação (Mestrado) – Mestrado em Ciência da Informação, Programa de Pós-graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis-SC, 2010.

MIORELLI, Sandra Teresinha. **Extração do sintagma nominal em sentenças em português.** 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) –

Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

MOENS, Marie-Francine. **Automatic indexing and abstracting of document texts**. [S.l.]: Springers, 2000. 284p.

MORELLATO, Luana Vieira. **Metodologia Computacional para Identificação de Sintagmas Nominais na Língua Portuguesa**. 2010. 112 f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2010.

_____. **SIDSN**: sistema identificador de sintagmas nominais, 2007. 58 f. Monografia (Bacharelado em Ciência da Computação) – Departamente de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2007.

NARUKAWA, C. M.; LEIVA, I. G.; FUJITA, M. N. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software sisa com uso da terminologia decs na área de odontologia. **Informação & Sociedade: Estudos**, v. 19, n. 2, p. 99-118, 2009. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/a/7566>. Acesso em: 01 Jan. 2017.

NASCIMENTO, Gustavo Diniz do. **Dos sintagmas nominais aos descritores documentais:** estudo de caso na indexação de teses e dissertações da área de direito. 2015, 198 f. Dissertação (Mestrado) – Mestrado em Ciência da Informação, Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2015.

NASCIMENTO, Gustavo Diniz do; CORRÊA, R. F. Sintagmas nominais com valor de descritores: critérios para seleção. **XVII ENANCIB**, v. 17, 2016.

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

_____. Interface entre lingüística e indexação: uma revisão de literatura. **Rev. Bras. Biblio. Doc**., São Paulo, v.21, n. 1/2, p. 46-62, jan./jun. 1988.

NAVES, Madalena M. L. **Curso de indexação**: princípios e técnicas de indexação, com vistas à recuperação da informação. Belo Horizonte: UFMG, Biblioteca Universitária, 2004. Material didático. 23p.

NEET, H. E. **L´analyse documentaire** : notes et documentation destinées aux étudiants de l´École de Bibliothècaires. Génève: Institut d´Études Sociales. École de Bibliotécaires, 1989.

NORONHA, Daisy Pires; FERREIRA, Sueli Mara S. P. Revisões de literatura. In: CAMPELLO, Bernadete Santos; CONDÓN, Beatriz Valadares; KREMER, Jeannette Marguerite (orgs.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: UFMG, 2000.

OLIVEIRA, Marlene de. (Coord.). **Ciência da informação e biblioteconomia**: novos conteúdos e espaços de atuação. Belo Horizonte: UFMG, 2011.

OTHERO, G. A. **A gramática da frase em português**: algumas reflexões para a formalização da estrutura frasal em português. Porto Alegre: EDIPUCRS, 2009.

PEREIRA, Edmeire Cristina; BUFREM, Leilah Santiago. Princípios de organização e representação de conceitos em linguagens documentarias. **Enc. BIBLI: R. Electrônica de Bibl. Ci. Inform.**, Florianópolis, n.20, 2° semestre 2005. Disponível em: <http://www.encontros-bibli.ufsc.br/sumario.htm> Acesso em: 15 ago. 2016

PERINI, M. A.; FRAIHA, S.; FULGÊNCIO, L.; NETO, R. B. O SN em português: A hipótese mórfica. Belo Horizonte: **Revista de Estudos de Linguagem** - UFMG, Jul./Dez. 1996. p.43-56

PERINI, M. A. **Gramática descritiva do português**. 3.ed. São Paulo: Ática, 1998

PERINI, M. A. **Gramática descritiva do português**. 4.ed. São Paulo: Ática, 2005

PERINI, MÁRIO A. **Gramática do português brasileiro**. São Paulo: Parábola, 2010.

PINTO, Virgínia Bentes; MEUNIER, Jean-Guy. **Les images visuelles**: un regard sur leur représentation indexale. Montreal: [s.n.]. 2006.

PINTO, Virgínia Bentes; MEUNIER, Jean-Guy; SILVA NETO, Casemiro. A contribuição peirciana para a representação indexal de imagens visuais. **Enc. BIBLI: R. Electrônica de Bibl. Ci. Inform**., Florianópolis, n. 25, p. 15-35, 1º sem.2008. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/view/1153/878> Acesso em: 15 ago. 2016

RAJU, J.; RAJU, R. **Descriptive and subject cataloguing**: a workbook. Oxford: Chandos Publishing, 2006.

REDIGOLO, Franciele Marques. **O processo de análise de assunto na catalogação de documentos**: a perspectiva sociocognitiva do catalogador em contexto de Biblioteca Universitária. 2010, 176 f. Dissertação (Mestrado em Ciência

da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, São Paulo, 2010.

RIBEIRO, C. L. M. NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (org.). **Organização da informação**: princípios e tendências. Brasília-DF: Briquet de Lemos, 2006. 142p. ISBN: 85-85637-30-7. **Revista Brasileira de Biblioteconomia e Documentação**, v. 2, n. 2, p. 104-106, 2006. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/a/6229>. Acesso em: 15 ago. 2016

ROBREDO, Jaime. Organização dos documentos ou organização da informação: uma questão de escolha. **DataGramaZero.,** v.5 n.1 fev/04.

_____. A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. O., ed. **Estudos Avançados em Biblioteconomia e Ciência da Informação**. Brasília, ABDF, 1982. v. 1, p. 236-74.

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

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

RUIZ PEREZ, R. **El analisis documental**: bases terminológicas, conceptualización y estructura operativa. Granada : Ed. Universidad de Granada, 1992.

SALTON, Gerard.; MCGILL, M.J. **Introduction to Modern Information Retrieval**. McGraw-Hill, New York, NY, 1983.

SALTON, Gerard. Recent studies in automatic text analysis and document retrieval. **Journal of the Association for Computing Machinery**, v. 20, n. 2, p. 258-27, Apr. 1973.

_____. Automatic text analysis: automatic document indexing and classification methods are examined and their effectiveness assessed. **Science**, 168 (3929):335-43, 17 Abr. 1970.

SÁNCHEZ LUNA, Blanca Estela. “Catalogación por materia”. In: Figueroa Alcántara, Hugo Alberto; Ramírez Velásquez, César Augusto. **Organización bibliográfica y documental**. México: CUIBUNAM, 2004, pp. 83-103. ISBN 970-32-1861-X.

SANTOS, Cícero Nogueira dos. **Aprendizado de máquina na identificação de sintagmas nominais:** o caso do português brasileiro. 2005. 104 f. Dissertação

(Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2005.

SARACEVIC, Tefko. Information Science. **JASIS – Journal of the American Society for Information Science**, v. 50, n. 12, p. 1051-1063, 1999.

_____. Ciência da informação: origem, evolução e relações. **Perspec. Ci. Inf.**, Belo Horizonte, v. 1, n. 1, p .41-62, jan./jun. 1996.

SAUPERL, A. **Subject determination during the catalog process**. Lanham: Scarecrow, 2002.

SAUTCHUK, Inez. **Prática de morfossintaxe**: como e por que aprender análise (morfo) sintática. 2.ed. – Barueri, SP: Manole, 2010.

SHERA, J. H.; EGAN, M. E. Exame atual da Biblioteconomia e da Documentação. In: BRADFORD, S. C. **Documentação**. Rio de Janeiro: Editôra Fundo de Cultura, 1961. p. 15- 61.

SILVA, C. V. da. **O contexto do catalogador de assuntos em bibliotecas universitárias**: aspectos físicos, psicológicos e sócio-cognitivos. 2006. 113 f. Trabalho de Conclusão de Curso (Graduação em Biblioteconomia) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2006.

SILVA, Maria R; FUJITA, Mariângela S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, v. 16, n. 2, p. 133-161, maio/ago. 2004.

SILVA, Tiago. Jose da.; CORRÊA, R. F. Ferramentas para indexação automática: uma análise comparativa entre o ogma, parser palavras, lx-parser e a extração manual de sintagmas nominais. **XVII Encontro Nacional de Pesquisa em Ciência da Informação**, v. 16, 2015.

SILVA, Tiago Jose da. **Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa**. 2014, 144 f. Dissertação (Mestrado) – Mestrado em Ciência da Informação, Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2014.

SOUZA, B. P. de.; FUJITA, M. S. L. Análise de assunto no processo de indexação: um percurso entre teoria e norma**. Inf. & Soc.:Est**., João Pessoa, v.24, n.1, p. 19-34, jan./abr. 2014

SOUZA, Renato Rocha. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. Tese (Doutorado em Ciência

da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

SOUZA, Renato Rocha.; RAGHAVAN, K. S. A methodology for noun phrase-based automatic indexing. **Knowledge Organization**, v. 33, n. 1, p. 45-56, 2006

SOUZA, Renato Rocha. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais**. Enc. BIBLI: R. Electrônica de Bibl. Ci. Inform.**, Florianópolis, v. 11, n. esp., p. 42-59, 1º sem. 2006.

SOUZA, Renato Rocha; RAGHAVAN, K. S. Extraction of keywords from texts: an exploratory study using Noun Phrases. **Informação & Tecnologia (ITEC)**. Marília/ João Pessoa. v. 1, n. 1. p. 5-16, jan./jun., 2014.

STREHL, Letícia. Avaliação da consistência da indexação realizada em uma biblioteca universitária de artes. **Ci. Inf.**, v. 27, n. 3, p. 329-335, set./dez. 1998.

UNISIST. Princípios de indexação. **Revista da Escola de Biblioteconomia da UFMG**. Belo Horizonte, v.10, n.1, p.83- -94, mar. 1981.

UNISIST. Princípios de indexação. Tradução de Maria Cristina M. F. Pinto. **Revista da Escola de Biblioteconomia da UFMG**., Belo Horizonte, v. 1, n. 10, p. 83-94, mar. 1981. Título original: Indexing Principles

UNESCO. **Guidelines for the establishment and development of monolingual thesauri**. Paris, 1973. 37 p

VALE, Eunides A. do. Linguagens de indexação. In: SMIT, Johanna W. (Coord.). **Análise documentária: a análise da síntese**. Brasília: NCT – CNPQ , 1987. Cap. 1, p.12 – 26.

VAN SLYPE, G. **Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales**. Trad. Pedro Hípola e Félix de Moya. Madrid: Fundación Germán Sánchez Ruipérez; Pirámide, 1991. 200p. Tradução de: Les languages d'indexation: conception, construction et utilisation dans les systèmes documentaires.

VIEIRA, Renata et al. Extração de sintagmas nominais para o processamento de coreferência. In: Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR), 5. **Anais**... São Carlos: ICMC/USP, 2000. p. 165-173.

VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Ci. Inf**. Brasília, 17 (1): 43-57, jan./jun. 1988.

VOGEL, Michely Jabala Mamede. **A noção de estrutura linguística e de processo de estruturação e sua influência no conceito e na elaboração de linguagens**

documentárias. 2007, 137 f. Dissertação (Mestrado em Ciência da Informação). Universidade de São Paulo. 2007.

WIVES, Leandro K. **Indexação de documentos textuais**. 1997. 19f. Trabalho Monográfico - Disciplina de Sistemas de Banco de Dados (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997.

APÊNDICE A – DOCUMENTOS E PALAVRAS-CHAVE

O Apêndice A é responsável por apresentar a referência e as palavras-chave dos documentos utilizados no *corpus* da pesquisa. O seu objetivo é facilitar a apresentação de determinados dados da pesquisa e facilitar que o leitor localize esses documentos caso seja necessário.

DOC 01

RF: SANTOS, Plácida Leopoldina Ventura Amorim da Costa; SANT´ANA, Ricardo César Gonçalves. Transferência da informação: análise para valoração de unidades de conhecimento. **DataGramaZero**, Rio de Janeiro, v. 3, n. 2, abr. 2002.
KW: Transferência de informação; Gestão do conhecimento; Valor de unidades de conhecimento.

DOC 02

RF: MUELLER, Suzana Pinheiro Machado. Popularização do conhecimento científico. **DataGramaZero**, Rio de Janeiro, v. 3, n. 2, abr. 2002.
KW: Popularização da Ciência; Comunicação Científica.

DOC 03

RF: CASTRO, Ana Lúcia Siaines de. O valor da informação: um desafio permanente. **DataGramaZero**, Rio de Janeiro, v. 3, n. 3, jun. 2002.
KW: Informação; Valor Informacional; Direito à Informação; Memória Social; Estoque Informacional.

DOC 04

RF: CAFÉ, Lígia; LAGE, Márcia Basílio. Auto-arquivamento: uma opção inovadora para a produção científica. **DataGramaZero**, Rio de Janeiro, v. 3, n. 3, jun. 2002.
KW: Arquivos-abertos; Sistema de Publicação; Budapest Open Access Initiative; Acesso Livre; Auto-arquivamento.

DOC 05

RF: BURNHAM, Teresinha Fróes. Análise Contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público. **DataGramaZero**, Rio de Janeiro, v. 3, n. 3, jun. 2002.
KW: Conhecimento Científico; Conhecimento Privado; Conhecimento Escolar; Democratização da Ciência; Comunicação Científica.

DOC 06

RF: LEVACOV, Marília; VANTI, Nadia; ZANCAN, Júlio César; MENDES, Maria Lizete Gomes. O Tesauro Eletrônico do Mundo do Trabalho: produto de um esforço interdisciplinar. **DataGramaZero**, Rio de Janeiro, v. 3, n. 4, ago. 2002.
KW: Tesauro Eletrônico; Mundo do Trabalho; Recuperação da Informação; Interface de Consulta; Sistema de Informação; Interdisciplinaridade; Interação Humano-Computador (IHC).

DOC 07

RF: VALENTIM, Marta Lígia Pomim. Inteligência competitiva em organizações: dado, informação e conhecimento. **DataGramaZero**, Rio de Janeiro, v. 3, n. 4, ago. 2002.
KW: Inteligência Competitiva; Gestão do Conhecimento; Gestão da Informação; Fluxos Informacionais; Transferência da Informação.

DOC 08

RF: MIRANDA, A. L. C.; SIMEÃO, Elmira. A conceituação de massa documental e o ciclo de interação entre tecnologia e o registro do conhecimento. **DataGramaZero**, Rio de Janeiro, v. 3, n. 4, ago. 2002.
KW: Informação; Massa Documental; Conceito de Informação; Tecnologia; Registro do Conhecimento.

DOC 09

RF: STAREC, Claudio. Informação e universidade: os pecados informacionais e barreiras na comunicação da informação para a tomada de decisão na universidade. **DataGramaZero**, Rio de Janeiro, v. 3, n. 4, ago. 2002.
KW: Universidade; Gestão do fluxo de Informação na Universidade; Inteligência Competitiva; Barreiras na Comunicação da Informação; Pecados Informacionais.

DOC 10

RF: PORCARO, Rosa Maria. Implicações da ´nova economia´ para a mensuração estatística: desajustes conceituais e metodológicos.**DataGramaZero**, Rio de Janeiro, v. 3, n. 4, ago. 2002.
KW: Informação Estatística; Nova Economia; Mensuração Estatística; Desajuste Conceitual; Metodologia Estatística.

DOC 11

RF: PATERNOSTRO, Luiz Carlos Brito. Por uma nova Ciência da Informação: ensino, pesquisa e formação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.
KW: Ciência da Informação; Armazenamento e recuperação; Curso em informação; Unidade e especificidade da informação.

DOC 12

RF: DIAS, Eduardo José Wense. Ensino e pesquisa em Ciência da informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.
KW: Ciência da Informação; Biblioteconomia; Sistema de Informação; Arquivologia; Ensino; Pesquisa.

DOC 13

RF: CARVALHO, Kátia de. O profissional da informação: o humano multifacetado. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.
KW: Profissional da informação; Informação organizacional; Formação e profissional da informação.

DOC 14

RF: TARAPANOFF, Kira; SUAIDEN, Emir José; OLIVEIRA, Cecília Leite. Funções sociais e oportunidades para profissionais da informação.**DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.

KW: Profissionais da informação; Funções sociais; Perfis de profissionais da informação; Inclusão digital; Gestão da informação; Gestão do conhecimento.

DOC 15

RF: RODRIGUES, Mara Eliane Fonseca. Relação ensino-pesquisa: em discussão a formação do profissional da informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.
KW: Formação profissional; Ensino e pesquisa.

DOC 16

RF: CARDOSO, Ana Maria P.. Educação para a informação: desafios contemporâneos para a Ciência da Informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, out. 2002.
KW: Ciência da Informação - Formação profissional; Educação Superior no Brasil; Sociedade da Informação - educação; Ciência da Informação e Biblioteconomia; Ciência da Informação - curso de graduação.

DOC 17

RF: TARGINO, Maria das Graças. Novas tecnologias e produção científica: uma relação de causa e efeito ou uma relação de muitos efeitos?.**DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Internet e Produção Científica; Novas Tecnologias de Informação e de Comunicação; Produção Científica e Novas Tecnologias.

DOC 18

RF: DAGNINO, Renato Peixoto. Enfoques sobre a relação ciência, tecnologia e sociedade: neutralidade e determinismo. **DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Estudos Sociais da Ciência; Sociologia da Ciência; Ciência e Sociedade; Tecnologia e Sociedade.

DOC 19

RF: BARBOSA, Ricardo Rodrigues. Inteligência empresarial: uma avaliação de fontes de informação sobre o ambiente organizacional externo.**DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Inteligência Empresarial; Monitoração Ambiental; Fontes de Informação; Gestão do Conhecimento; Gestão da Informação

DOC 20

RF: SMIT, Johanna W.; DIAS, Eduardo José Wense; SOUZA, Rosali Fernandez de. Contribuição da pós-graduação para a Ciência da Informação no Brasil: uma visão. **DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Ciência da Informação no Brasil; Avaliação 2001 CAPES; Pós-graduação em Ciência da Informação; Pesquisa em Ciência da Informação no Brasil.

DOC 21

RF: DUMONT, Lígia Maria Moreira. Os múltiplos aspectos e interfaces da leitura. **DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Leitura-teoria; Cognóscio; Conhecimento-introjeção; Leitura e Sociedade; Informação e Sociedade.

DOC 22

RF: SANTOS, Nilton Bahlis dos. A Informação e o Paradigma Holográfico: a Utopia de Vannevar Bush. **DataGramaZero**, Rio de Janeiro, v. 3, n. 6, dez. 2002.
KW: Paradigma; Holografia; Ciência da Informação; Tecnologia da Informação; Hipertexto; Complexidade; Interatividade; Virtual; Totalidade.

DOC 23

RF: COSTA, Icléia Thiesen Magalhães. Informação, Memória e Espaço Prisional no Rio de Janeiro. **DataGramaZero**, Rio de Janeiro, v. 4, n. 1, fev. 2003.
KW: Informação; Memória Social; Espaço Prisional.

DOC 24

RF: GONZÁLEZ DE GÓMEZ, Maria Nélida. O contrato social da pesquisa: em busca de uma nova equação entre a autonomia epistêmica e autonomia política. **DataGramaZero**, Rio de Janeiro, v. 4, n. 1, fev. 2003.
KW: Contrato Social; Ciência; Pesquisa; Pesquisadores; Autonomia; Ecologia dos Conhecimentos.

DOC 25

RF: MUELLER, Suzana Pinheiro Machado; SANTANA, Maria Gorete Henrique. A Ciência da Informação no CNPq - fomento à formação de recursos humanos e à pesquisa entre 1994-2002. **DataGramaZero**, Rio de Janeiro, v. 4, n. 1, fev. 2003.
KW: Fomento à pesquisa - Ciência da Informação; CNPq - fomento á pesquisa em Ciência da Informação.

DOC 26

RF: BARRETO, Aldo de Albuquerque. Políticas de monitoramento da informação por compressão semântica dos seus estoques.**DataGramaZero**, Rio de Janeiro, v. 4, n. 2, abr. 2003.
KW: Compressão Semântica; Monitoramento da Informação; Estoques de Informação; Palavras-chave.

DOC 27

RF: OLINTO, Gilda. Bolsas de Pesquisador do CNPq: informações sobre política de C&T a partir da base que contém os dados cadastrais dos bolsistas. **DataGramaZero**, Rio de Janeiro, v. 4, n. 2, abr. 2003.
KW: Indicadores Científicos; Política Científica e Tecnológica; Gestão de Ciência e Tecnologia

DOC 28

RF: PACHECO, Roberto Carlos dos Santos; KERN, Vinícius Medina. Arquitetura conceitual e resultados da integração de sistemas de informação e gestão da ciência e tecnologia. **DataGramaZero**, Rio de Janeiro, v. 4, n. 2, abr. 2003.
KW: Governo Eletrônico; Arquitetura de Sistemas de Informação; Integração de Informações; Gestão de C&T; Bibliotecas Digitais; Plataforma Lattes; Rede ScienTI.

DOC 29

RF: MARCONDES, Carlos Henrique; JARDIM, José Maria. Políticas de informação governamental: a construção de governo eletrônico na administração federal do Brasil. **DataGramaZero**, Rio de Janeiro, v. 4, n. 2, abr. 2003.
KW: Governo Eletrônico; Políticas de Informação; Informação Governamental.

DOC 30

RF: DIAS, Guilherme Ataíde. Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 7-12, jan./abr. 2002.
KW: Periódicos eletrônicos; Avaliação de acesso; Arquivo de log de acesso.

DOC 31

RF: GONZÁLEZ DE GÓMEZ, Maria Nélida. Novos cenários políticos para a informação. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 27-40, jan./abr. 2002.
KW: Política de informação; Sociedade da informação; Internet; Institucionalização da informação; Estado.

DOC 32

RF: LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.
KW: Bases de dados; Estratégia de busca; Linguagem controlada; Linguagem natural. Recuperação da informação; Artigo de revisão.

DOC 33

RF: OHIRA, Maria Lourdes Blatt; PRADO, Noêmia Schoffen. Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000).**Ciência da Informação**, Brasília, v. 31, n. 1, p. 61-74, jan./abr. 2002.
KW: Biblioteca digital; Biblioteca virtual; Produção científica: Produção bibliográfica; Periódicos.

DOC 34

RF: PRYSTHON, Cecília F.; SCHMIDT, Susana. Experiência do Leaal/UFPE na produção e transferência de tecnologia. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 84-90, jan./abr. 2002.
KW: Informação tecnológica; Transferência de informação; Transferência tecnológica.

DOC 35

RF: ALMEIDA, Maurício Barcellos. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 5-13, maio/ago. 2002.
KW: XML; HTML; Linguagens de marcação; Internet; Intranet.

DOC 36

RF: URBIZAGÁSTEGUI ALVARADO, Rubén. A Lei de Lotka na bibliometria brasileira. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 14-20, maio/ago. 2002.
KW: Bibliometria; Lei de Lotka; Produtividade de autores; Brasil.

DOC 37

RF: CENDÓN, Beatriz Valadares. Bases de dados de informação para negócios. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 30-43, maio/ago. 2002.
KW: Informação para negócios; Bases de dados

DOC 38

RF: GARCEZ, Eliane Maria Stuart; RADOS, Gregório Jean Varvakis. Biblioteca híbrida: um novo enfoque no suporte à educação a distância.**Ciência da Informação**, Brasília, v. 31, n. 2, p. 44-51, maio/ago. 2002.
KW: Biblioteca híbrida; Tipos de usuários; Bens e serviços.

DOC 39

RF: LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002.
KW: Estratégia de busca; Recuperação da informação; Técnicas de estratégia de busca; Bases de dados; Artigo de revisão.

DOC 40

RF: MARCHIORI, Patrícia Zeni. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 72-79, maio/ago. 2002.
KW: Ciência da informação; Gestão da informação.

DOC 41

RF: POBLACIÓN, Dinah Aguiar; NORONHA, Daisy Pires. Produção das literaturas “branca” e “cinzenta” pelos docentes/doutores dos programas de pós-graduação em ciência da informação no Brasil. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 98-106, maio/ago. 2002.
KW: Produção científica; Literatura branca; Literatura cinzenta; Ciência da informação.

DOC 42

RF: REZENDE, Yara. Informação para negócios: os novos agentes do conhecimento e a gestão do capital intelectual. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 120-128, maio/ago. 2002.
KW: Gestão do conhecimento; Capital intelectual; Informação para negócios; Sistemas de informação para negócios; Agentes do conhecimento.

DOC 43

RF: SILVA, Janete Fernandes; FERREIRA, Marta Araújo Tavares; BORGES, Mônica Erichsen Nassif. Análise metodológica dos estudos de necessidades de informação sobre setores industriais brasileiros: proposições. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 129-141, maio/ago. 2002.
KW: Necessidade de informação tecnológica; Informação tecnológica; Setor industrial; Inovação.

DOC 44

RF: SILVA, Sergio Luis da. Informação e competitividade: a contextualização da gestão do conhecimento nos processos organizacionais. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 142-151, maio/ago. 2002.

KW: Gestão do conhecimento; Informação e competitividade; Processos organizacionais.

DOC 45

RF: VANTI, Nadia. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 152-162, maio/ago. 2002.
KW: Bibliometria; Cienciometria; Informetria; Webometria; Métodos quantitativos de avaliação

DOC 46

RF: BORGES, Paulo César Rodrigues. Métodos quantitativos de apoio à bibliometria: a pesquisa operacional pode ser uma alternativa?. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 5-17, set./dez. 2002.
KW: Bibliometria; Lei de Bradford; Pesquisa operacional; Caos; Ciência da informação; Inferência bayesiana.

DOC 47

RF: DIAS, Guilherme Ataíde. Periódicos eletrônicos: considerações relativas à aceitação deste recurso pelos usuários. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 18-25, set./dez. 2002.
KW: Periódicos eletrônicos; Usabilidade; Novas tecnologias.

DOC 48

RF: COHEN, Max F.. Alguns aspectos do uso da informação na economia da informação. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 26-36, set./dez. 2002.
KW: Uso da informação; Economia da Informação; Modelo genérico.

DOC 49

RF: ORTIZ, Lúcia Cunha; ORTIZ, Wilson Aires; SILVA, Sergio Luis da. Ferramentas alternativas para monitoramento e mapeamento automatizado do conhecimento. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 66-76, set./dez. 2002.
KW: Monitoramento da informação; Biblioteconomia; Ciência da informação.

DOC 50

RF: SILVA, Edna Lúcia da; CUNHA, Miriam Vieira da. A formação profissional no século XXI: desafios e dilemas. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 77-82, set./dez. 2002.
KW: Educação dos bibliotecários; Profissional da informação.

DOC 51

RF: TORRES, Elisabeth Fátima; MAZZONI, Alberto Angel; ALVES, João Bosco da Mota. A acessibilidade à informação no espaço digital. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 83-91, set./dez. 2002.
KW: Acessibilidade; Espaço digital; Bibliotecas; Pessoas portadoras de deficiência; Ajudas técnicas.

DOC 52

RF: FREIRE, Isa Maria; NATHANSOHN, Bruno Macedo; TAVARES, Carla; ESPÍRITO SANTO, Carmelita do. Estudos de usuários: o padrão que une três abordagens. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 103-107, set./dez. 2002.
KW: Estudos de usuários; Educação ambiental; Internet; Hipertexto; Pesquisa participante.

DOC 53

RF: AIRES, Rachel Virgínia Xavier; ALUÍSIO, Sandra Maria. Como incrementar a qualidade dos resultados das máquinas de busca: da análise de logs à interação em português. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 5-16, jan./abr. 2003.
KW: Análise de logs; Máquinas de busca; Recuperação de informação; Comportamento de usuários; Estratégias de busca.

DOC 54

RF: DUDZIAK, Elisabeth Adriana. Information literacy: princípios, filosofia e prática. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 23-35, jan./abr. 2003.
KW: Information literacy; Competência em informação; Alfabetização informacional; Biblioteca aprendente; Bibliotecário educador; Sociedade de aprendizagem; Habilidades informacionais.

DOC 55

RF: FERREIRA, Danielle Thiago. Profissional da informação: perfil de habilidades demandadas pelo mercado de trabalho. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 42-49, jan./abr. 2003.
KW: Profissional da informação; Profissional da informação – habilidades; Perfil e atuação profissional; Mercado de trabalho.

DOC 56

RF: FREIRE, Isa Maria. O olhar da consciência possível sobre o campo científico. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 50-59, jan./abr. 2003.
KW: Teoria da ciência da informação; Sociologia da informação; História da ciência da informação; Comunicação científica; Responsabilidade social.

DOC 57

RF: GONZÁLEZ DE GÓMEZ, Maria Nélida. As relações entre ciência, Estado e sociedade: um domínio de visibilidade para as questões da informação. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 60-76, jan./abr. 2003.
KW: Recuperação da informação; Inteligência científica; Integração dos conhecimentos; Estado; Ciência; Sociedade; Informação.

DOC 58

RF: LIMA, Gercina Ângela Borém de Oliveira. Interfaces entre a ciência da informação e a ciência cognitiva. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 77-87, jan./abr. 2003.
KW: Ciência da informação; Ciência cognitiva; Processamento da informação; Categorização; Indexação; Recuperação da informação; Interação homem-computador.

DOC 59

RF: MOSTAFA, Solange Puntel; MÁXIMO, Luis Fernando. A produção científica da Anped e da Intercom no GT da Educação e Comunicação.**Ciência da Informação**, Brasília, v. 32, n. 1, p. 96-101, jan./abr. 2003.
KW: Comunicação científica; Bibliometria; Comunicação e Educação; Estudo de citações; Cientometria.

DOC 60

RF: SILVA, Helena Pereira da. Inteligência competitiva na Internet: um processo otimizado por agentes inteligentes. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 115-134, jan./abr. 2003.
KW: Inteligência competitiva; Internet; Monitoramento de fontes de informação; Agentes inteligentes.

APÊNDICE B – LISTA DE SINTAGMAS NOMINAIS DIVIDIDOS NA ETAPA 1

O Apêndice B é responsável por apresentar os SNs que tiveram alterações devido a aplicação da etapa 1. Nele é possível identificar qual SN teve alteração (negrito e sublinhado), qual o critério da etapa 1 que foi utilizado para essa alteração e quais SNs resultaram dessa alteração.

Doc.	Critério	Sintagma Nominal
Art. 01	Preposição + artigo	**[os estudos] ~~sobre a~~ [gestão do conhecimento]**
		os estudos
		A gestão do conhecimento
Art. 02	Pela	**[comunicação científica] ~~pela~~ [ciência da informação]**
		comunicação científica
		A ciência da informação
Art. 02	Pela	**[estudos da comunicação científica] pela [ciência da informação]**
		estudos da comunicação científica
		A ciência da informação
Art. 02	Pela	**[interesse para estudos da comunicação científica] pela [ciência da informação]**
		interesse para estudos da comunicação científica
		A ciência da informação
Art. 02	(1) Expressão Comum (artigo/ preposição + Tema de); (2) Pela	**um tema de [interesse para estudos da comunicação científica] pela [ciência da informação]**
		Interesse para estudos da comunicação científica
		A ciência da informação
Art. 03	Que + Verbo	**[um estoque informacional de valor social] que possibilita**
		Um estoque informacional de valor social
Art. 04	Parênteses	**a Budapest Open Access Initiative (BOAI)**
		a Budapest Open Access Initiative
Art. 04	Parênteses	**A experiência da Budapest Open Access Initiative (BOAi)**
		A experiência da Budapest Open Access Initiative
Art. 04	Parênteses	**O acesso livre (access)**
		O acesso livre
Art. 04	(1) Conjunção + Pronome Possessivo	**[o conceito inovador de auto-arquivamento] e suas [implicações no sistema de publicações científicas]**
		o conceito inovador de auto-arquivamento
		Implicações no sistema de publicações científicas
Art. 04	(1) Preposição + Verbo; (2) Que + Verbo	**[o objetivo] de mostrar [uma ação efetiva] que viabiliza [o auto-arquivamento]**

		o objetivo
		uma ação efetiva
		o auto-arquivamento
Art. 04	Que + Verbo	**[uma ação efetiva] que viabiliza [o auto-arquivamento]**
		uma ação efetiva
		o auto-arquivamento
Art. 05	Preposição + artigo	**[A dinâmica de construção] de uma [metodologia de análise de processos de tradução do conhecimento científico]**
		A dinâmica de construção
		Metodologia de análise de processos de tradução do conhecimento científico
Art. 05	Preposição + artigo	**[construção] de uma [metodologia de análise de processos de tradução do conhecimento científico]**
		construção
		Metodologia de análise de processos de tradução do conhecimento científico
Art. 06	Expressão Comum (Artigo + Ponto + De + Vista + De)	**o ponto de vista da [Interação Humano-Computador]**
		A Interação Humano-Computador
Art. 07	Que + Verbo + Preposição	**[ações] que contribuem para [o desenvolvimento da inteligência competitiva organizacional]**
		ações
		O desenvolvimento da inteligência competitiva organizacional
Artigo 08	(1) Preposição (Entre); (2) Conjunção + Artigo	**[interação] entre [tecnologia] e o [registro do conhecimento]**
		interação
		tecnologia
		registro do conhecimento
Art. 08	(1) Preposição (Entre); (2) Conjunção + Artigo	**[o ciclo de interação] entre [tecnologia] e o [registro do conhecimento]**
		o ciclo de interação
		tecnologia
		registro do conhecimento
Art. 09	Expressão Comum (Artigo + Foco + De)	**O foco da [Inteligência Competitiva]**
		A Inteligência Competitiva
Art. 10	Expressão Comum (Implicações + De)	**Implicações da ["nova economia"]**
		A nova economia
Art. 11	(1) Expressão Comum (Capaz + De + Verbo)	**[Ciência da Informação] capaz de tratar**
		Ciência da Informação
Art. 11	(1) Expressão Comum (Capaz + De + Verbo)	**[um curso de Ciência da Informação] capaz de tratar**
		um curso de Ciência da Informação
Art. 14	Conjunção + Verbo (Participio)	**[funções sociais] e delineados**
		Funções Sociais

Art. 18	(1) Artigo + forma(s) + de + Verbo; (2) Artigo + Campo + De	**[as] ~~formas de abordar~~ [o] ~~campo dos~~ [Estudos Sociais da Ciência e Tecnologia ou]**
		Os Estudos Sociais da Ciência e Tecnologia ou
Art. 18	Artigo + Campo + De + Artigo	[o] ~~campo dos~~ [Estudos Sociais da Ciência e Tecnologia ou]
		Os Estudos Sociais da Ciência e Tecnologia ou
Art. 19	(1) Preposição + Que + Verbo; (2) Expressão Comum (Diversos Tipos de)	**[a freqüência] com que utilizam diversos tipos de [fontes de informação]**
		a freqüência
		fontes de informação
Art. 19	Expressão Comum (Diversos Tipos de)	**diversos tipos de [fontes de informação]**
		fontes de informação
Art. 19	Expressão Comum (Sobre o)	**[fontes de informação] sobre o [ambiente organizacional externo]**
		fontes de informação
		ambiente organizacional externo
Art. 19	Expressão Comum (Sobre o)	**[uma avaliação de fontes de informação] sobre o [ambiente organizacional externo]**
		uma avaliação de fontes de informação
		ambiente organizacional externo
Art. 20	(1) Pela; (2) Parênteses	**[Síntese da avaliação continuada dos programas de pós-graduação em Ciência da Informação reconhecidos] pela [CAPES] (PUC/CAMPUFBAUFMGUFRJ/IBICTUnB e UNESP/Marília)**
		Síntese da avaliação continuada dos programas de pós-graduação em Ciência da Informação reconhecidos
		A CAPES
Art. 22	(1) Expressão Comum (Artigo + Própria (o))	**a própria [definição de Ciência da Informação]**
		definição de Ciência da Informação
Art. 22	Preposição (Deste)	**[no interior] deste [paradigma]**
		no interior
		paradigma
Art. 22	Artigo + Problema + De	**o problema da [complexidade e interatividade]**
		A complexidade e interatividade
Art. 22	(1) Expressão Comum (Artigo + Própria(o))	**o próprio [paradigma vigente]**
		paradigma vigente
Art. 22	(1) Expressão Comum (Possibilidade de); (2) Verbo + Preposição; (3) Artigo + Problema + Preposição	**possibilidade da [Ciência da Informação] enfrentar de uma [maneira nova] o problema da [complexidade e interatividade]**
		A Ciência da Informação
		Maneira nova
		A complexidade e interatividade
Art. 23	Expressão Comum (Relação entre)	**[As] relações entre [informação]**
		A informação

Art. 23	Expressão Comum (Contida Em)	**[da informação histórica] contida nos [escaninhos da memória]**
		da informação histórica
		Os escaninhos da memória
Art. 23	Expressão Comum (Contida Em)	**[disseminação da informação histórica] contida nos [escaninhos da memória]**
		disseminação da informação histórica
		Os escaninhos da memória
Art. 24	Expressão Comum (Possibilidade de)	**[A] possibilidade de [reformulação do contrato social da ciência]**
		A reformulação do contrato social da ciência
Art. 24	Expressão Comum (definição de); (2) Conjunção + Preposição; (3) Que + Verbo	**[as] definições dos [sujeitos] e dos [princípios] que organizam [os programas de pesquisa]**
		os sujeitos
		os princípios
		Os programas de pesquisa
Art. 24	(1) Expressão Comum (Busca de); (2) Preposição (Entre) + Artigo	**busca de [uma nova equação] entre a [autonomia epistêmica e autonomia política]**
		uma nova equação
		autonomia epistêmica e autonomia política
Art. 24	Que + Verbo	**[dos princípios] que organizam [os programas de pesquisa]**
		dos princípios
		os programas de pesquisa
Art. 24	(1) Conjunção + Preposição; (2) Que + Verbo	**[dos sujeitos] e dos [princípios] que organizam [os programas de pesquisa]**
		dos sujeitos
		os princípios
		Os programas de pesquisa
Art. 24	(1) Preposição (Entre) + Artigo	[uma nova equação] entre a [autonomia epistêmica e autonomia política]
		uma nova equação
		autonomia epistêmica e autonomia política
Art. 25	Expressão Comum (Após a); (2) Conjunção + Pronome Possessivo	**Após [as origens do CNPq] é sua [vocação inicial]**
		as origens do CNPq
		vocação inicial
Art. 25	(1) Após + Artigo/Preposição; (2) Conjunção + Pronome Possessivo	**Após do [CNPq] é sua [vocação inicial]**
		O CNPq
		vocação inicial
Art. 26	Preposição + Pronome Possessivo	**[Compressão Semântica] dos seus [Estoques]**
		Compressão Semântica
		Estoques

Art. 27	(1) Expressão Comum (Finalidade de); (2) Verbo	[a] finalidade de gerar [indicadores científicos e tecnológicos]
		indicadores científicos e tecnológicos
Art. 27	(1) Que + Pronome Pessoal + Verbo; (2) Expressão Comum (Através de); (3) Preposição (Dessa); (4) Expressão Comum (Gerado(a) + Preposição); (5) Preposição + Artigo (Para uma); (6) Preposição + Artigo (com a) + (7) Expressão Comum (Finalidade de); (8) Verbo	**[análises] que se apresentam através da [transformação] dessa [base de dados] gerada com [fins administrativos] para uma base com a finalidade de gerar [indicadores científicos e tecnológicos]**
		análises
		A transformação
		A base de dados
		fins administrativos
		indicadores científicos e tecnológicos
Art. 27	(1) Expressão Comum (Possibilidad de); (2) Que + Pronome Pessoal + Verbo; (3) Expressão Comum (Através de); (4) Preposição (Dessa); (5) Expressão Comum (Gerado(a) + Preposição); (6) Preposição + Artigo (Para uma); (7) Preposição + Artigo (com a) + (8) Expressão Comum (Finalidade de); (9) Verbo	**possibilidades de [análises] que se apresentam através da [transformação] dessa [base de dados] gerada com [fins administrativos] para uma base com a finalidade de gerar [indicadores científicos e tecnológicos]**
		análises
		A transformação
		A base de dados
		fins administrativos
		indicadores científicos e tecnológicos
Art. 28	Expressão Comum (Iniciativa de)	**iniciativas de [governo eletrônico]**
		Governo eletrônico
Art. 28	(1) Expressão Comum (Artigo + Papel + Preposição)	~~**O papel das**~~ **[bibliotecas digitais de teses e dissertações]**
		As bibliotecas digitais de teses e dissertações
Art. 29	Expressão Comum (Artigo + Noção + Preposição)	**[a] noção de [governo eletrônico]**
		Governo eletrônico
Art. 29	Expressão Comum (Artigo+ Impacto + Preposição)	**o impacto do [Governo Eletrônico]**
		O Governo Eletrônico
Art. 29	Expressão Comum (Obstáculo + Artigo/Preposição)	**Outro obstáculo ao [Governo Eletrônico]**
		O Governo Eletrônico
Art. 30	(1) Preposição (Pela)	~~**pela**~~ **[análise do arquivo de log de acesso]**
		A análise do arquivo de log de acesso
Art. 31	(1) Expressão Comum (Artigo + Papel + Preposição)	**[as mudanças]** ~~**do papel do**~~ **[Estado]**
		as mudanças
		O Estado
Art. 31	Expressão Comum (O Papel De)	~~**do papel do**~~ **[Estado]**

		O Estado
Art. 32	(1) Expressão Comum (Artigo + Função + Preposiçao); (2) Preposição + Preposição; (3) Expressão Comum (Nesse Contexto)	**a função do [vocabulário controlado] ou da [linguagem natural] nesse contexto**
		O vocabulário controlado
		A linguagem natural
Art. 32	Preposição + Preposição	**[da linguagem controlada] ou da [linguagem natural]**
		da linguagem controlada
		A linguagem natural
Art. 32	Expressão Comum (Nesse Contexto)	**[da linguagem natural] nesse contexto**
		da linguagem natural
Art. 32	Expressão Comum (Preposição/Artigo + Planejamento + Preposição)	**do planejamento da [estratégia de busca]**
		A estratégia de busca
Art. 32	(1) Preposição + Preposição; (2) Expressão Comum (Nesse Contexto)	**[do vocabulário controlado] ou da [linguagem natural] nesse contexto**
		do vocabulário controlado
		A linguagem natural
Art. 32	Preposição + Preposição	**[o uso da linguagem controlada] ou da [linguagem natural]**
		o uso da linguagem controlada
		A linguagem natural
Art. 33	(2) Verbo + Preposição	**[A evolução da temática biblioteca virtual e biblioteca digital como assunto de artigos de periódicos brasileiros] publicados de [1995 a 2000]**
		A evolução da temática biblioteca virtual e biblioteca digital como assunto de artigos de periódicos brasileiros
		1995 a 2000
Art. 35	(1) Expressão Comum (A Utilização De); (2) Parênteses	**a utilização do [Extended Markup Language] (XML)**
		O Extended Markup Language
Art. 35	Preposição + Verbo	**[alguns conceitos complementares necessários ao entendimento do assunto] em apresentar [vantagens no uso do XML]**
		Alguns conceitos complementares necessários ao entendimento do assunto
		Vantagens no uso do XML
Art. 35	Preposição + Verbo	**[ao entendimento do assunto] ~~em apresentar~~ [vantagens no uso do XML]**
		Ao entendimento do assunto
		Vantagens no uso do XML
Art. 35	Expressão Comum (Cada Tipo De)	**cada tipo de [navegador da Internet]**
		Navegador da Internet
Art. 35	Preposição + Verbo	**[do assunto] em apresentar [vantagens no uso do XML]**
		do assunto

		Vantagens no uso do XML
Art. 35	Parênteses	**[do Extended Markup Language] (XML)**
		Do Extended Markup Language
Art. 35	(1) Que + Verbo + Advérbio; (2) Preposição + Que	**[páginas Html] que possuem mais [marcações] do que [conteúdo]**
		páginas Html
		marcações
		conteúdo
Art. 37	Expressão comum (Das Principais)	**das principais [bases de dados estrangeiras sobre informação]**
		bases de dados estrangeiras sobre informação
Art. 37	(1) Expressão comum (Também + Algum (as) (uns)); Expressão comum (Das Principais)	**também algumas das principais [empresas produtoras e distribuidoras de bases de dados sobre informação da indústria de informação eletrônica]**
		empresas produtoras e distribuidoras de bases de dados sobre informação da indústria de informação eletrônica
Art. 38	(1) Expressão comum (A Importância da); (2) Que + Verbos + Preposição; (3) Preposição + Verbo; (4) Preposição + Artigo	**a importância da [flexibilização dos bens e serviços] que devem ser oferecidos pelas [bibliotecas híbridas] para atender [às necessidades] de uma [diversidade de tipos de usuários existentes na educação a distância]**
		A flexibilização dos bens e serviços
		As bibliotecas híbridas
		às necessidades
		A diversidade de tipos de usuários existentes na educação a distância
Art. 38	Preposição + Artigo	**[às necessidades] de uma [diversidade de tipos de usuários existentes na educação a distância]**
		às necessidades
		A diversidade de tipos de usuários existentes na educação a distância
Art. 38	(1) Que + Verbos + Preposição; (2) Preposição + Verbo; (3) Preposição + Artigo	**[flexibilização dos bens e serviços] que devem ser oferecidos pelas [bibliotecas híbridas] para atender [às necessidades] de uma [diversidade de tipos de usuários existentes na educação a distância]**
		flexibilização dos bens e serviços
		As bibliotecas híbridas
		às necessidades
		A diversidade de tipos de usuários existentes na educação a distância
Art. 38	(1) Pelas; (2) Preposição + Verbo; (3) Preposição + Artigo	**pelas [bibliotecas híbridas] para atender [às necessidades] de uma [diversidade de tipos de usuários existentes na educação a distância]**
		bibliotecas híbridas
		às necessidades

		A diversidade de tipos de usuários existentes na educação a distância
Art. 38	(1) Que + Verbos + Preposição; (2) Preposição + Verbo; (3) Preposição + Artigo	**[serviços] que devem ser oferecidos pelas [bibliotecas híbridas] para atender [às necessidades] de uma [diversidade de tipos de usuários existentes na educação a distância]**
		serviços
		As bibliotecas híbridas
		às necessidades
		A diversidade de tipos de usuários existentes na educação a distância
Art. 40	Expressão Comum (Artigo + Área + Preposição)	**da área de [ciência da informação]**
		A ciência da informação
Art. 40	(1) Expressão Comum (Os Pressupostos Teóricos); (2) Expressão Comum (A área da)	**os pressupostos teóricos da área de [ciência da informação]**
		ciência da informação
Art. 40	(1) Verbo + Que; (2) Preposição + Pronome Demonstrativo	**Ressalta-se que [a gestão da informação compartilha] com demais [profissões afins]**
		A gestão da informação compartilha
		Profissões afins
Art. 41	Expressão Comum (A área da)	**[545 % titulados] na área da [ciência da informação]**
		545 % titulados
		A ciência da informação
Art. 41	Parênteses	**[da literatura cinzenta produzida] (402%)**
		da literatura cinzenta produzida
Art. 41	Expressão Comum (A análise dos)	**na análise dos [5 Programas em Ciência da Informação]**
		Os 5 Programas em Ciência da Informação
Art. 41	Expressão Comum (Na Área da)	**na área da [ciência da informação]**
		A ciência da informação
Art. 42	Conjunção + Preposição	**[da administração do conhecimento] e do [capital intelectual das empresas]**
		da administração do conhecimento
		O capital intelectual das empresas
Art. 42	Conjunção + Preposição	**[da importância estratégica da administração do conhecimento] e do [capital intelectual das empresas]**
		da importância estratégica da administração do conhecimento
		O capital intelectual das empresas
Art. 42	Conjunção + Preposição	**[do conhecimento] e do [capital intelectual das empresas]**
		do conhecimento
		O capital intelectual das empresas
Art. 42	(1) Expressão Comum (Artigo + Reconhecimento + Preposição); (2) Conjunção + Preposição	**O reconhecimento da [importância estratégica da administração do conhecimento] e do [capital intelectual das empresas]**

		A importância estratégica da administração do conhecimento
		O capital intelectual das empresas
Art. 43	(1) Expressão Comum (As Futuras Pesquisas); (2) Expressão Comum (Sobre); (3) Verbo + Pelo(s); (4) Conjunção + Preposição	**as futuras pesquisas sobre [necessidades informacionais] ditadas pelos [processos de aprendizagem] e da [inovação tecnológica]**
		Necessidades informacionais
		Os processos de aprendizagem
		a inovação tecnológica
Art. 43	Verbo + Preposição (EM)	**[informação tecnológica] detectados em [empresas brasileiras do setor industrial]**
		informação tecnológica
		Empresas brasileiras do setor industrial
Art. 43	Verbo + Preposição (EM)	**[necessidade de informação tecnológica] detectados em [empresas brasileiras do setor industrial]**
		necessidade de informação tecnológica
		Empresas brasileiras do setor industrial
Art. 43	(1) Verbo + Pelo(s); (2) Conjunção + Preposição	**[necessidades informacionais] ditadas pelos [processos de aprendizagem] e da [inovação tecnológica]**
		necessidades informacionais
		Os processos de aprendizagem
		a inovação tecnológica
Art. 43	Verbo + Preposição (EM)	**[os diagnósticos de necessidade de informação tecnológica] detectados em [empresas brasileiras do setor industrial]**
		Os diagnósticos de necessidade de informação tecnológica
		Empresas brasileiras do setor industrial
Art. 43	Conjunção + Preposição	**[pelos processos de aprendizagem] e da [inovação tecnológica]**
		pelos processos de aprendizagem
		a inovação tecnológica
Art. 43	(1) Sobre + Artigo; (2) Verbo + Preposição (EM)	**[uma investigação] sobre os [diagnósticos de necessidade de informação tecnológica] detectados em [empresas brasileiras do setor industrial]**
		uma investigação
		Os diagnósticos de necessidade de informação tecnológica
		Empresas brasileiras do setor industrial
Art. 44	Expressão Comum (Ligado + a + Artigo)	**[da gestão do conhecimento] ligado à [aprendizagem]**
		da gestão do conhecimento
		A aprendizagem
Art. 44	Expressão Comum (Ligado + a + Artigo)	**[o lado operacional da gestão do conhecimento] ligado à [aprendizagem]**
		o lado operacional da gestão do conhecimento
		A aprendizagem

Art. 46	(1) Expressão Comum (Alguns pontos); (2) Que + Verbo; (3) Adjetivo Isolado; (4) Entre + Artigo; (5) Conjunção + Artigo	**alguns pontos que parecem [comuns] entre a [bibliometria] e a [Teoria do Caos]**
		Comuns
		a bibliometria
		a Teoria do Caos
Art. 46	(1) Travessão; (2) Verbo + Pela (o); (3) Parênteses	**[analogia entre fenômenos físicos da Teoria do Caos] – resolvidos pela [Pesquisa Operacional] (PO)**
		Analogia entre fenômenos físicos da Teoria do Caos
		a pesquisa operacional
Art. 46	Expressão Comum (As várias)	**as várias [formulações no campo da bibliometria]**
		Formulações no campo da bibliometria
Art. 46	(1) Travessão; (2) Verbo + Pela (o); (3) Parênteses	**[fenômenos físicos da Teoria do Caos] – resolvidos pela [Pesquisa Operacional] (PO)**
		Fenômenos físicos da Teoria do Caos
		a pesquisa operacional
Art. 46	(1) Pela; (2) Parênteses	pela [Pesquisa Operacional] (Po)
		a pesquisa operacional
Art. 46	(1) Expressão Comum (Uma + Linha + De); (2) Travessão; (3) Verbo + Pela (o); (4) Parênteses	**uma linha de [analogia entre fenômenos físicos da Teoria do Caos] – resolvidos pela [Pesquisa Operacional] (PO)**
		Analogia entre fenômenos físicos da Teoria do Caos
		a pesquisa operacional
Art. 46	(1) Expressão Comum (Uma Saída Para); (2) Preposição + Verbo	**uma saída para sistematizar [conceitos na bibliometria]**
		conceitos na bibliometria
Art. 47	(1) Sobre + Artigo	**[algumas reflexões] sobre a [aceitação de periódicos eletrônicos disponibilizados]**
		algumas reflexões
		aceitação de periódicos eletrônicos disponibilizados
Art. 48	(1) Que + Verbos; (2) Expressão Comum (Por + Parte + De)	**[a construção do modelo] que permita medir o [uso da informação] por parte [das organizações]**
		a construção do modelo
		uso da informação
		das organizações
Art. 48	(1) Que + Verbos; (2) Expressão Comum (Por + Parte + De)	**[do modelo] que permita medir o [uso da informação] por parte [das organizações]**
		O modelo
		uso da informação
		das organizações
Art. 50	Expressão Comum (Enfoque Especial)	**enfoque especial [à educação dos bibliotecários]**
		à educação dos bibliotecários
Art. 50	(1) Expressão Comum (Enfoque Especial)	**[no século XXI com] enfoque especial [à educação dos bibliotecários]**
		no século XXI com

		à educação dos bibliotecários
Art. 51	(1) Expressão Comum (Relacionado (a) + Artigo); (2) Preposição + Artigo	**[às situações] relacionadas à [interação das pessoas portadoras de deficiência] com a [informação]**
		às situações
		interação das pessoas portadoras de deficiência
		informação
Art. 51	(1) Preposição + Verbo; (2) Preposição + Artigo	**[o intuito] de contribuir para um [maior nível de acessibilidade à informação]**
		o intuito
		maior nível de acessibilidade à informação
Art. 53	(1) Que + Verbos	**[informações] que pudessem facilitar a [recuperação de informação]**
		informações
		recuperação de informação
Art. 54	(1) Preposição + Artigo; (2) Expressão Comum (Ante a); (3) Que + Verbo	**[a necessidade de construção] de um [novo paradigma educacional] ante a [sociedade atual] que incorpore [a competência em informação]**
		a necessidade de construção
		novo paradigma educacional
		sociedade atual
		a competência em informação
Art. 54	Expressão Comum (A Própria)	**a própria [essência da competência em informação]**
		Essência da competência em informação
Art. 54	Que + Verbo	**[a sociedade atual] que incorpore [a competência em informação]**
		A sociedade atual
		a competência em informação
Art. 54	(1) Preposição + Artigo; (2) Expressão Comum (Ante a); (3) Que + Verbo	**[construção] de um [novo paradigma educacional] ante a [sociedade atual] que incorpore [a competência em informação]**
		Construção
		novo paradigma educacional
		sociedade atual
		a competência em informação
Art. 54	(1) Expressão Comum (Ante a); (2) Que + Verbo	**[Um novo paradigma educacional] ante a [sociedade atual] que incorpore [a competência em informação]**
		Um novo paradigma educacional
		sociedade atual
		a competência em informação
Art. 55	Expressão Comum (A + Demanda + Atual + De)	**da demanda atual do [mercado de trabalho]**
		O mercado de trabalho
Art. 55	Verbo + Pelo(s)	**[habilidades] demandadas pelo [mercado de trabalho]**
		habilidades
		O mercado de trabalho

Art. 55	Verbo + Pelo(s)	**[perfil de habilidades] demandadas pelo [mercado de trabalho]**
		perfil de habilidades
		o mercado de trabalho
Art. 56	Preposição + Artigo	**[a proposição] de uma [responsabilidade social]**
		a proposição
		responsabilidade social
Art. 57	Expressão Comum (A + Origem + De)	**a origem da [ciência da informação]**
		A ciência da informação
Art. 57	Expressão Comum (A Pesquisa Em)	**a pesquisa em [questões da informação]**
		Questões da informação
Art. 58	(1) Expressão Comum (Aspectos da); (2) Parênteses	**aspectos da [ciência da informação] (CI)**
		A ciência da informação
Art. 58	Parênteses	**[da ciência cognitiva] (CC)**
		da ciência cognitiva
Art. 58	Parênteses	**[da ciência da informação] (CI)**
		Da ciência da informação
Art. 58	(1) Sobre + Expressão Comum; (2) Parênteses	**[Estudo panorâmico] sobre aspectos da [ciência da informação] (CI)**
		Estudo panorâmico
		A ciência da informação
Art. 58	(1) Entre + Artigo; (2) Conjunção + Artigo	**[Interfaces] entre a [ciência da informação] e a [ciência cognitiva]**
		Interfaces
		Ciência da informação
		Ciência cognitiva
Art. 60	Expressão Comum (Por meio de)	[a automação do processo] por meio de [agentes inteligentes]
		a automação do processo
		agentes inteligentes
Art. 60	(1) Expressão Comum (Proposta + De); (2) Parênteses	**[a] proposta de [um processo de inteligência competitiva] (IC)**
		Um processo de inteligência competitiva
Art. 60	Expressão Comum (Por meio de)	**[do processo] por meio de [agentes inteligentes]**
		do processo
		agentes inteligentes
Art. 60	Parênteses	**[inteligência competitiva] (IC)**
		inteligência competitiva
Art. 60	Expressão Comum (A + Tarefa + de)	**na tarefa de [monitoramento de fontes de informação disponíveis na rede]**
		Monitoramento de fontes de informação disponíveis na rede
Art. 60	Parênteses	**um processo de inteligência competitiva (IC)**
		um processo de inteligência competitiva

Art. 60	Verbo + Preposição (POR)	**[um processo] otimizado por [agentes inteligentes]**
		um processo
		agentes inteligentes

APÊNDICE C – REMOÇÃO DE ARTIGOS DO INICIO E FIM DOS SNS

No Apêndice C são apresentados os dados relativos ao critério de remoção de artigos do início e fim dos SNs, aplicados durante a etapa 2 da pesquisa. Nesse apêndice é apresentado uma tabela com os SNs antes e depois da aplicação do critério e a qual documento cada SN representa.

Documento	Sintagma Nominal Antes	Sintagma Nominal Depois
Artigo 01	a gestão do conhecimento	gestão do conhecimento
Artigo 01	os estudos	estudos
Artigo 01	a gestão do conhecimento	gestão do conhecimento
Artigo 02	A questão da popularização da ciência	questão da popularização da ciência
Artigo 02	A ciência da informação	ciência da informação
Artigo 02	A ciência da informação	ciência da informação
Artigo 02	A ciência da informação	ciência da informação
Artigo 02	A ciência da informação	ciência da informação
Artigo 03	A análise da informação	análise da informação
Artigo 03	a informação	informação
Artigo 03	a informação	informação
Artigo 03	a informação	informação
Artigo 03	a informação	informação
Artigo 03	a questão da informação	questão da informação
Artigo 03	O Valor da Informação	Valor da Informação
Artigo 03	Um estoque informacional de valor social	estoque informacional de valor social
Artigo 04	a Budapest Open Access Initiative	Budapest Open Access Initiative
Artigo 04	A experiência da Budapest Open Access Initiative	experiência da Budapest Open Access Initiative
Artigo 04	O acesso livre	acesso livre
Artigo 04	o auto-arquivamento	auto-arquivamento
Artigo 04	o conceito inovador de auto-arquivamento	conceito inovador de auto-arquivamento
Artigo 04	o objetivo	objetivo
Artigo 04	uma ação efetiva	ação efetiva
Artigo 04	o auto-arquivamento	auto-arquivamento
Artigo 04	uma ação efetiva	ação efetiva
Artigo 04	o auto-arquivamento	auto-arquivamento
Artigo 05	A dinâmica de construção	dinâmica de construção
Artigo 05	a tradução de conhecimento científico em conhecimento público	tradução de conhecimento científico em conhecimento público
Artigo 05	o conhecimento científico	conhecimento científico
Artigo 05	uma metodologia de análise de processos de tradução do conhecimento científico	metodologia de análise de processos de tradução do conhecimento científico
Artigo 06	a Interação Humano-Computador	Interação Humano-Computador
Artigo 06	o gerenciamento do Tesauro Eletrônico do Mundo do Trabalho	gerenciamento do Tesauro Eletrônico do Mundo do Trabalho

Artigo 06	A Interação Humano-Computador	Interação Humano-Computador
Artigo 06	O Tesauro Eletrônico do Mundo do Trabalho	Tesauro Eletrônico do Mundo do Trabalho
Artigo 06	O Tesauro Eletrônico do Mundo do Trabalho	Tesauro Eletrônico do Mundo do Trabalho
Artigo 07	A consolidação do processo de inteligência competitiva organizacional	consolidação do processo de inteligência competitiva organizacional
Artigo 07	A inteligência competitiva organizacional	inteligência competitiva organizacional
Artigo 07	O desenvolvimento da inteligência competitiva organizacional	desenvolvimento da inteligência competitiva organizacional
Artigo 07	O processo de inteligência competitiva organizacional	processo de inteligência competitiva organizacional
Artigo 07	O processo de inteligência competitiva	processo de inteligência competitiva
Artigo 08	A Ciência da Informação	Ciência da Informação
Artigo 08	A conceituação de massa documental	conceituação de massa documental
Artigo 08	A polissemia do conceito de informação	polissemia do conceito de informação
Artigo 08	o ciclo de interação	ciclo de interação
Artigo 08	O conceito de informação	conceito de informação
Artigo 08	O registro do conhecimento	registro do conhecimento
Artigo 09	a Inteligência Competitiva	Inteligência Competitiva
Artigo 09	a tomada de decisão na universidade	tomada de decisão na universidade
Artigo 09	A Universidade Estácio de Sá	Universidade Estácio de Sá
Artigo 09	os gestores da universidade	gestores da universidade
Artigo 09	os pecados informacionais	pecados informacionais
Artigo 10	a mensuração estatística	mensuração estatística
Artigo 10	A nova economia	nova economia
Artigo 11	um curso de Ciência da Informação	curso de Ciência da Informação
Artigo 12	a pesquisa	pesquisa
Artigo 12	a questão básica da ciência da informação	questão básica da ciência da informação
Artigo 13	O Profissional da Informação	Profissional da Informação
Artigo 14	as funções sociais delineadas	funções sociais delineadas
Artigo 16	a Ciência da Informação	Ciência da Informação
Artigo 18	Os Estudos Sociais da Ciência e Tecnologia ou	Estudos Sociais da Ciência e Tecnologia ou
Artigo 18	Os Estudos Sociais da Ciência e Tecnologia ou	Estudos Sociais da Ciência e Tecnologia ou
Artigo 19	a freqüência	freqüência
Artigo 19	uma avaliação de fontes de informação	avaliação de fontes de informação
Artigo 20	A CAPES	CAPES
Artigo 22	Ciência da Informação	Ciência da Informação
Artigo 22	O Paradigma Holográfico	Paradigma Holográfico
Artigo 22	O Paradigma Holográfico	Paradigma Holográfico
Artigo 22	A complexidade e interatividade	complexidade e interatividade
Artigo 22	A Ciência da Informação	Ciência da Informação
Artigo 22	A complexidade e interatividade	complexidade e interatividade
Artigo 23	A Ciência da Informação	Ciência da Informação
Artigo 23	A informação	informação
Artigo 23	Os escaninhos da memória	escaninhos da memória
Artigo 23	Os escaninhos da memória	escaninhos da memória

Artigo 24	A autonomia epistêmica	autonomia epistêmica
Artigo 24	A reformulação do contrato social da ciência	reformulação do contrato social da ciência
Artigo 24	Os sujeitos	sujeitos
Artigo 24	os princípios	princípios
Artigo 24	Os programas de pesquisa	programas de pesquisa
Artigo 24	uma nova equação	nova equação
Artigo 24	os programas de pesquisa	programas de pesquisa
Artigo 24	os princípios	princípios
Artigo 24	Os programas de pesquisa	programas de pesquisa
Artigo 24	O Contrato Social da Pesquisa	Contrato Social da Pesquisa
Artigo 24	os programas de pesquisa	programas de pesquisa
Artigo 24	uma nova equação	nova equação
Artigo 25	as origens do CNPq	origens do CNPq
Artigo 25	O CNPq	CNPq
Artigo 25	As ações do CNPq relatadas	ações do CNPq relatadas
Artigo 27	A transformação	transformação
Artigo 27	A base de dados	base de dados
Artigo 27	A transformação	transformação
Artigo 27	A base de dados	base de dados
Artigo 28	A internacionalização da Plataforma Lattes	internacionalização da Plataforma Lattes
Artigo 28	A Plataforma Lattes	Plataforma Lattes
Artigo 28	As bibliotecas digitais de teses e dissertações	bibliotecas digitais de teses e dissertações
Artigo 29	A construção de Governo Eletrônico	construção de Governo Eletrônico
Artigo 29	O Governo Eletrônico	Governo Eletrônico
Artigo 29	O Governo Eletrônico	Governo Eletrônico
Artigo 30	à análise do arquivo de log de acesso	análise do arquivo de log de acesso
Artigo 30	a avaliação do acesso a periódicos eletrônicos disponibilizados	avaliação do acesso a periódicos eletrônicos disponibilizados
Artigo 30	As características inerentes à análise do arquivo de log de acesso	características inerentes à análise do arquivo de log de acesso
Artigo 30	O arquivo de log de acesso da revista Informação & Sociedade	arquivo de log de acesso da revista Informação & Sociedade
Artigo 30	A análise do arquivo de log de acesso	análise do arquivo de log de acesso
Artigo 31	as mudanças	mudanças
Artigo 31	O Estado	Estado
Artigo 31	O Estado	Estado
Artigo 32	O vocabulário controlado	vocabulário controlado
Artigo 32	A linguagem natural	linguagem natural
Artigo 32	A linguagem natural	linguagem natural
Artigo 32	A estratégia de busca	estratégia de busca
Artigo 32	A linguagem natural	linguagem natural
Artigo 32	o uso da linguagem controlada	uso da linguagem controlada
Artigo 32	A linguagem natural	linguagem natural
Artigo 32	um ambiente de bases de dados	ambiente de bases de dados

Artigo 33	A evolução da temática biblioteca virtual e biblioteca digital como assunto de artigos de periódicos brasileiros	evolução da temática biblioteca virtual e biblioteca digital como assunto de artigos de periódicos brasileiros
Artigo 34	os mecanismos de transferência tecnológica no âmbito da universidade e comunidade externa	mecanismos de transferência tecnológica no âmbito da universidade e comunidade externa
Artigo 35	A Internet	Internet
Artigo 35	O Extended Markup Language	Extended Markup Language
Artigo 35	a versão 40 do Html	versão 40 do Html
Artigo 35	O Html	Html
Artigo 35	o uso intensivo do Html	uso intensivo do Html
Artigo 35	O XML	XML
Artigo 35	Uma introdução ao XML	introdução ao XML
Artigo 36	A Lei de Lotka	Lei de Lotka
Artigo 36	as aplicações da Lei de Lotka	aplicações da Lei de Lotka
Artigo 38	A flexibilização dos bens e serviços	flexibilização dos bens e serviços
Artigo 38	As bibliotecas híbridas	bibliotecas híbridas
Artigo 38	às necessidades	necessidades
Artigo 38	A diversidade de tipos de usuários existentes na educação a distância	diversidade de tipos de usuários existentes na educação a distância
Artigo 38	às necessidades	necessidades
Artigo 38	A diversidade de tipos de usuários existentes na educação a distância	diversidade de tipos de usuários existentes na educação a distância
Artigo 38	As bibliotecas híbridas	bibliotecas híbridas
Artigo 38	às necessidades	necessidades
Artigo 38	A diversidade de tipos de usuários existentes na educação a distância	diversidade de tipos de usuários existentes na educação a distância
Artigo 38	às necessidades	necessidades
Artigo 38	A diversidade de tipos de usuários existentes na educação a distância	diversidade de tipos de usuários existentes na educação a distância
Artigo 38	As bibliotecas híbridas	bibliotecas híbridas
Artigo 38	às necessidades	necessidades
Artigo 38	A diversidade de tipos de usuários existentes na educação a distância	diversidade de tipos de usuários existentes na educação a distância
Artigo 38	uma diversidade de tipos de usuários existentes na educação	diversidade de tipos de usuários existentes na educação
Artigo 40	A gestão da informação	gestão da informação
Artigo 40	A gestão da informação	gestão da informação
Artigo 40	A ciência da informação	ciência da informação
Artigo 40	A gestão da informação compartilha	gestão da informação compartilha
Artigo 41	A ciência da informação	ciência da informação
Artigo 41	A produção científica vinculada à linha de pesquisa	produção científica vinculada à linha de pesquisa
Artigo 41	a respectiva produção científica	respectiva produção científica
Artigo 41	Os 5 Programas em Ciência da Informação	5 Programas em Ciência da Informação
Artigo 41	A ciência da informação	ciência da informação
Artigo 42	a gestão do capital intelectual	gestão do capital intelectual
Artigo 42	O capital intelectual das empresas	capital intelectual das empresas
Artigo 42	O capital intelectual das empresas	capital intelectual das empresas
Artigo 42	O capital intelectual das empresas	capital intelectual das empresas

Artigo 42	A importância estratégica da administração do conhecimento	importância estratégica da administração do conhecimento
Artigo 42	O capital intelectual das empresas	capital intelectual das empresas
Artigo 42	Os novos agentes do conhecimento	novos agentes do conhecimento
Artigo 43	Os processos de aprendizagem	processos de aprendizagem
Artigo 43	A inovação tecnológica	inovação tecnológica
Artigo 43	Os processos de aprendizagem	processos de aprendizagem
Artigo 43	A inovação tecnológica	inovação tecnológica
Artigo 43	Os diagnósticos de necessidade de informação tecnológica	diagnósticos de necessidade de informação tecnológica
Artigo 43	A inovação tecnológica	inovação tecnológica
Artigo 43	uma investigação	investigação
Artigo 43	Os diagnósticos de necessidade de informação tecnológica	diagnósticos de necessidade de informação tecnológica
Artigo 44	a contextualização da gestão do conhecimento nos processos organizacionais	contextualização da gestão do conhecimento nos processos organizacionais
Artigo 44	a visualização da gestão do conhecimento na organização	visualização da gestão do conhecimento na organização
Artigo 44	A aprendizagem	aprendizagem
Artigo 44	o lado operacional da gestão do conhecimento	lado operacional da gestão do conhecimento
Artigo 44	A aprendizagem	aprendizagem
Artigo 45	a bibliometria	bibliometria
Artigo 45	caracterização da webometria	caracterização da webometria
Artigo 45	a cienciometria	cienciometria
Artigo 45	à informetria	informetria
Artigo 45	A webometria	webometria
Artigo 46	à bibliometria	bibliometria
Artigo 46	a bibliometria	bibliometria
Artigo 46	a ciência da informação	ciência da informação
Artigo 46	a pesquisa operacional	pesquisa operacional
Artigo 46	a Teoria do Caos	Teoria do Caos
Artigo 46	a bibliometria	bibliometria
Artigo 46	a Teoria do Caos	Teoria do Caos
Artigo 46	a pesquisa operacional	pesquisa operacional
Artigo 46	a pesquisa operacional	pesquisa operacional
Artigo 46	os métodos da Pesquisa Operacional	métodos da Pesquisa Operacional
Artigo 46	a pesquisa operacional	pesquisa operacional
Artigo 46	a pesquisa operacional	pesquisa operacional
Artigo 47	a aceitação de periódicos eletrônicos disponibilizados	aceitação de periódicos eletrônicos disponibilizados
Artigo 48	a construção do modelo	construção do modelo
Artigo 48	o modelo	modelo
Artigo 48	o uso da informação	uso da informação
Artigo 48	uma economia da informação	economia da informação
Artigo 50	à educação dos bibliotecários	educação dos bibliotecários
Artigo 50	à educação dos bibliotecários	educação dos bibliotecários
Artigo 50	à educação dos bibliotecários	educação dos bibliotecários

Artigo 51	A acessibilidade à informação no espaço digital	acessibilidade à informação no espaço digital
Artigo 51	a acessibilidade ao espaço digital	acessibilidade ao espaço digital
Artigo 51	à acessibilidade no espaço digital	acessibilidade no espaço digital
Artigo 51	à informação no espaço digital	informação no espaço digital
Artigo 51	à interação das pessoas portadoras de deficiência	interação das pessoas portadoras de deficiência
Artigo 51	às situações	situações
Artigo 51	o intuito	intuito
Artigo 51	um maior nível de acessibilidade à informação	maior nível de acessibilidade à informação
Artigo 52	a educação ambiental	educação ambiental
Artigo 52	uma experiência de interatividade na rede Internet	experiência de interatividade na rede Internet
Artigo 53	a qualidade dos resultados das máquinas de busca	qualidade dos resultados das máquinas de busca
Artigo 53	a recuperação de informação	recuperação de informação
Artigo 54	a necessidade de construção	necessidade de construção
Artigo 54	a competência em informação	competência em informação
Artigo 54	a sociedade atual	sociedade atual
Artigo 54	a competência em informação	competência em informação
Artigo 54	a competência em informação	competência em informação
Artigo 54	Um novo paradigma educacional	novo paradigma educacional
Artigo 54	A competência em informação	competência em informação
Artigo 55	O mercado de trabalho	mercado de trabalho
Artigo 55	O mercado de trabalho	mercado de trabalho
Artigo 55	O mercado de trabalho	mercado de trabalho
Artigo 55	o mercado de trabalho	mercado de trabalho
Artigo 56	a proposição	proposição
Artigo 56	uma responsabilidade social	responsabilidade social
Artigo 57	A ciência da informação	ciência da informação
Artigo 57	as questões da informação	questões da informação
Artigo 57	as relações entre ciência	relações entre ciência
Artigo 58	A ciência cognitiva	ciência cognitiva
Artigo 58	a ciência da informação	ciência da informação
Artigo 58	a ciência da informação	ciência da informação
Artigo 58	a ciência da informação	ciência da informação
Artigo 60	a automação do processo	automação do processo
Artigo 60	Um processo de inteligência competitiva	processo de inteligência competitiva
Artigo 60	um processo de inteligência competitiva	processo de inteligência competitiva
Artigo 60	um processo	processo

APÊNDICE D – REMOÇÃO DE PRONOMES

No Apêndice D são apresentados os dados relativos ao critério de remoção de pronomes, aplicados durante a etapa 2 da pesquisa. Nesse apêndice é apresentado uma tabela com os SNs antes e depois da aplicação do critério e a qual documento cada SN representa.

Documento	Sintagma Nominal Antes	Sintagma Nominal Depois
Artigo 22	deste paradigma	paradigma
Artigo 35	Alguns conceitos complementares necessários ao entendimento do assunto	conceitos complementares necessários ao entendimento do assunto
Artigo 35	sua utilização na Internet	utilização na Internet
Artigo 35	sua utilização na Internet	utilização na Internet
Artigo 37	muitas bases de dados	bases de dados
Artigo 47	algumas reflexões	Reflexões

APÊNDICE E – REMOÇÃO DE PREPOSIÇÃO E CONJUNÇÃO

No Apêndice E são apresentados os dados relativos ao critério de remoção de preposição e conjunção do início e fim dos SNs, aplicados durante a etapa 2 da pesquisa. Nesse apêndice é apresentado uma tabela com os SNs antes e depois da aplicação do critério e a qual documento cada SN representa.

Documento	Sintagma Nominal Antes	Sintagma Nominal Depois
Artigo 09	da universidade	universidade
Artigo 09	Na universidade	universidade
Artigo 10	da "nova economia"	nova economia
Artigo 12	da ciência da informação	ciência da informação
Artigo 18	dos Estudos Sociais da Ciência	Estudos Sociais da Ciência
Artigo 18	Estudos Sociais da Ciência e Tecnologia ou	Estudos Sociais da Ciência e Tecnologia
Artigo 18	Estudos Sociais da Ciência e Tecnologia ou	Estudos Sociais da Ciência e Tecnologia
Artigo 21	no cognóscio do leitor	cognóscio do leitor
Artigo 22	da Ciência da Informação	Ciência da Informação
Artigo 22	da complexidade	complexidade
Artigo 22	no interior	interior
Artigo 22	no paradigma do moderno	paradigma do moderno
Artigo 23	da informação histórica	informação histórica
Artigo 23	do espaço prisional	espaço prisional
Artigo 24	da atividade de pesquisa	atividade de pesquisa
Artigo 24	da ciência	ciência
Artigo 24	do contrato social da ciência	contrato social da ciência
Artigo 24	dos princípios	princípios
Artigo 24	dos sujeitos	sujeitos
Artigo 25	do CNPq	CNPq
Artigo 25	no CNPq	CNPq
Artigo 28	da Plataforma Lattes	Plataforma Lattes
Artigo 28	das bibliotecas digitais de teses e dissertações	bibliotecas digitais de teses e dissertações
Artigo 29	ao Governo Eletrônico	Governo Eletrônico
Artigo 29	do Governo Eletrônico	Governo Eletrônico
Artigo 30	da análise do arquivo de log de acesso	análise do arquivo de log de acesso
Artigo 30	do acesso periódicos eletrônicos disponibilizados	acesso periódicos eletrônicos disponibilizados
Artigo 30	do acesso periódicos eletrônicos	acesso periódicos eletrônicos
Artigo 30	do arquivo de log de acesso	arquivo de log de acesso
Artigo 30	do arquivo de log de acesso	arquivo de log de acesso
Artigo 30	do arquivo de log de acesso	arquivo de log de acesso
Artigo 31	do Estado	Estado
Artigo 32	da estratégia de busca	estratégia de busca

Artigo 32	da linguagem controlada	linguagem controlada
Artigo 32	da linguagem natural	linguagem natural
Artigo 32	da linguagem natural	linguagem natural
Artigo 32	do vocabulário controlado	vocabulário controlado
Artigo 33	da produção bibliográfica sobre bibliotecas virtuais e digitais	produção bibliográfica sobre bibliotecas virtuais e digitais
Artigo 33	da temática biblioteca virtual	temática biblioteca virtual
Artigo 34	da informação tecnológica	informação tecnológica
Artigo 35	Ao entendimento do assunto	entendimento do assunto
Artigo 35	ao Html	Html
Artigo 35	ao Html	Html
Artigo 35	ao XML	XML
Artigo 35	da Internet	Internet
Artigo 35	do assunto	assunto
Artigo 35	Do Extended Markup Language	Extended Markup Language
Artigo 35	do Html	Html
Artigo 35	do Html	Html
Artigo 35	do XML	XML
Artigo 35	na Internet	Internet
Artigo 35	na Internet	Internet
Artigo 35	no uso do XML	uso do XML
Artigo 36	da Lei de Lotka	Lei de Lotka
Artigo 36	na bibliometria brasileira	bibliometria brasileira
Artigo 39	na recuperação da informação	recuperação da informação
Artigo 41	da ciência da informação	ciência da informação
Artigo 41	da literatura cinzenta produzida	literatura cinzenta produzida
Artigo 41	dos Programas em Ciência da Informação	Programas em Ciência da Informação
Artigo 41	dos programas de pós-graduação em ciência da informação no Brasil	programas de pós-graduação em ciência da informação no Brasil
Artigo 42	da administração do conhecimento	administração do conhecimento
Artigo 42	da importância estratégica da administração do conhecimento	importância estratégica da administração do conhecimento
Artigo 42	do capital intelectual	capital intelectual
Artigo 42	do capital intelectual das empresas	capital intelectual das empresas
Artigo 42	do conhecimento	conhecimento
Artigo 43	da inovação tecnológica	inovação tecnológica
Artigo 43	do setor industrial	setor industrial
Artigo 43	pelos processos de aprendizagem	processos de aprendizagem
Artigo 44	da gestão do conhecimento	gestão do conhecimento
Artigo 44	da gestão do conhecimento na organização	gestão do conhecimento na organização
Artigo 44	da gestão do conhecimento processos organizacionais	gestão do conhecimento processos organizacionais
Artigo 44	do conhecimento processos organizacionais	conhecimento processos organizacionais
Artigo 44	nos processos organizacionais	processos organizacionais
Artigo 45	da bibliometria	bibliometria
Artigo 45	da webometria	webometria

Artigo 46	da bibliometria	bibliometria
Artigo 46	da Pesquisa Operacional	Pesquisa Operacional
Artigo 46	da Teoria do Caos	Teoria do Caos
Artigo 46	na bibliometria	bibliometria
Artigo 46	no campo da bibliometria	campo da bibliometria
Artigo 48	das organizações	organizações
Artigo 48	da informação na economia da informação	informação na economia da informação
Artigo 48	das organizações	organizações
Artigo 48	do uso da informação na economia da informação	uso da informação na economia da informação
Artigo 48	na economia da informação	economia da informação
Artigo 48	na economia da informação	economia da informação
Artigo 50	no século com	século
Artigo 51	ao espaço digital	espaço digital
Artigo 51	das pessoas portadoras de deficiência	pessoas portadoras de deficiência
Artigo 51	no espaço digital	espaço digital
Artigo 51	no espaço digital	espaço digital
Artigo 52	do hipertexto	hipertexto
Artigo 52	na estrutura do hipertexto	estrutura do hipertexto
Artigo 52	na rede Internet	rede Internet
Artigo 53	da análise de logs	análise de logs
Artigo 53	das máquinas de busca	máquinas de busca
Artigo 53	dos resultados das máquinas de busca	resultados das máquinas de busca
Artigo 54	da competência em informação	competência em informação
Artigo 55	do mercado de trabalho	mercado de trabalho
Artigo 55	pelo mercado de trabalho	mercado de trabalho
Artigo 55	pelo mercado de trabalho	mercado de trabalho
Artigo 57	da ciência da informação	ciência da informação
Artigo 57	da informação	informação
Artigo 57	da informação	informação
Artigo 58	da ciência cognitiva	ciência cognitiva
Artigo 58	da ciência cognitiva	ciência cognitiva
Artigo 58	Da ciência da informação	ciência da informação
Artigo 60	do processo	processo
Artigo 60	na Internet	Internet
Artigo 60	na Internet	Internet

APÊNDICE F – SINTAGMAS NOMINAIS NORMALIZADOS

Devido a extensão da lista dos SNs normalizados, não foi possível disponibilizar todos os SNs de forma física nesse trabalho sem comprometer a sua compreensão e estrutura. Portanto, aqui é apresentado (divido em três quadros) exemplos dos SNs alterados por cada um dos critérios até a etapa final de remoção de sufixo. Para facilitar também a localização das palavras-chave dos SNS foram destacadas em negrito e sublinhado. Para visualizar os demais SNs e suas alteração, é recomendado entrar no seguinte link: https://sites.google.com/site/renatocorrea/logic/ApendiceF.xlsx

Doc.	SN	1º Critério	2º Critério	3º Critério
Art. 01	A **gestão do conhecimento**	**Gestão do conhecimento**	**Gestão do conhecimento**	**Gestão do conhecimento**
Art. 02	Estudos da **comunicação científica**	Estudos da **comunicação científica**	Estudos da **comunicação científica**	Estudos da **comunicação científica**
Art. 09	Da **universidade**	Da **universidade**	Da **universidade**	Da **universidade**
Art. 22	Deste **paradigma**	Deste **paradigma**	**Paradigma**	**Paradigma**
Art. 35	A versão 40 do **Html**	Versão 40 do **Html**	Versão 40 do **Html**	Versão 40 do **Html**

Doc.	4º Critério	5º Critério	6º Critério	7º Critério
Art. 01	**Gestão do conhecimento**	**Gestão do conhecimento**	**Gestão do conhecimento**	gesta conhec
Art. 02	Estudos da **comunicação científica**	Estudos da **comunicação científica**	Estudos da **comunicação científica**	estud comunic cientif
Art. 09	Da **universidade**	Da **universidade**	**Universidade**	Univers
Art. 22	**Paradigma**	**Paradigma**	**Paradigma**	Paradigm
Art. 35	Versão do **Html**	Versão do **Html**	Versão do **Html**	versa html

Doc.	Resultado Final	Idêntico	Não São Termos do TBCI	Contêm Termos do TBCI
Art. 01	**Gestão do conhecimento**	1		
Art. 02	Estudos da **comunicação científica**			1
Art. 09	**Universidade**	1		
Art. 22	**Paradigma**	1		
Art. 35	Versão do **Html**			1